DOCUMENT RESUME

ED 241 586 TM 840 133

AUTHOR Ludlow, Larry H.
TITLE On the Simulation and Analysis of Measurement Model Residuals.
PUB DATE 11 Jan 84
NOTE 68p.; A paper presented at the Educational Testing Service, January 11, 1984. Based on "The Analysis of Rasch Model Residuals," doctoral dissertation, University of Chicago, 1983.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Data Analysis; Evaluation Methods; Goodness of Fit; *Latent Trait Theory; *Models; *Research Methodology; Research Needs; Simulation; Statistical Analysis
IDENTIFIERS Data Interpretation; Measurement Problems; *Rasch Model; *Residuals (Statistics)

ABSTRACT

The purpose of this research is to demonstrate that a systematic approach to the graphical analysis of Rasch model residuals can lead to an increased understanding of ordered response data, and that residual patterns do change in predictable ways, and that summary statistics need not be the only piece of evidence for assuring the fit between model and data. Three simple, idealized simulations and then two sets of real data are considered. The research concludes that (1) the measurement error uncovered in the residual analyses was not noticeable in the examination of person and item estimates, nor the person and item fit statistics; and (2) the tailored residuals provided a specific frame of reference within which the observed variation would be understood. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED241586

On the Simulation and Analysis of Measurement Model Residuals

by

Larry H. Ludlow

Boston College

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☑ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L H. Ludlow

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

January 11, 1984

TM 840 133

On The Simulation and Analysis of Measurement Model Residuals

A presentation prepared for Educational Testing Service

by Larry H. Ludlow, Boston College

January 11, 1984

I became attracted to the analysis of measurement model residuals, specifically the Rasch model, because it was apparent that the practical techniques commonly applied to regression, anova, factor analytic and practically any other statistical model residual were not being applied in the area of measurement.

Techniques for the inspection of residuals had been proposed but there was no boundary defining graphical investigation of residual patterns for data which fit the model, nor for deviations from the model when specific forms of data misfit are encountered. There was nothing like Draper & Smith to turn to.

Whenever an analysis of fit was discussed it was usually in terms of summary fit statistics and, as summary statistics, they do not usually provide the detailed interaction information that I am interested in when I analyze a set of data. Thus, the purpose of this research was to demonstrate that a systematic approach to the graphical analysis of Rasch model residuals can lead to an increased understanding of ordered response data, and that residual patterns do alter in predictable ways, and that summary statistics need not be our only piece of evidence for assuring the fit between model and data.

First, we will consider 3 simple, idealized simulations and then look at two sets of real data.

To reveal deviation from a model requires a background against which deviations are apparent. A background can be provided through the analysis of residuals from data simulated to fit the model. This research concentrated on two approaches to generating simulation data. In the first, item and person parameters are sampled from specified distributions and then data are generated to fit the model, given those parameters. I refer to these as "random" simulations. This method is useful for exploring the effect

that certain test characteristics have upon the distribution of residuals.

In the second approach, I begin with the analysis of observed data. Then, the item and person estimates from those data are used as the simulation parameters to generate data to fit the model. These are referred to as "tailored" simulations. Residuals produced by this method provide the relevant framework for revealing deviations from the model in observed data. This is because identical test characteristics should produce residuals which behave similarly if the observed data fit the model. The random simulations, with generally relevant parameter distributions, establish the broad background for what might be expected. The tailored simulations, with the observed estimates as parameters, focus on the observed data and define its particular baseline.

The following 3 simulation studies utilized 100 persons, 20 items, 3 response categories, (scored 0,1,2) and twenty replications. These simulations use the Rating Scale model but the results hold for Dichotomous or Partial Credit data.

Study 1 investigates the distribution of residuals under the limiting condition that all person measures and item calibrations equal zero. The intent is to reveal the pattern of residual variation when the parameters of the model are limited to the least variation possible.

Figure 1 is a residual-by-measure plot for Study 1. Three clusters of residuals are apparent. If the data had been generated so that all measures and calibrations received estimates of exactly zero, then the expected score for every person on every item would be 1. Residuals, in that case, would be just three points on this plot. Responses of "2" would have a residual of 1, "1" responses would yield 0 valued residuals, and "0" responses would have yielded -1 valued residuals. When standardized, those values would be 0.0, and -1.2. Those three values lie in the center of each of these clusters.

The three clusters, and not just 3 points, result because the generated data produced estimates that only approximated not equalled the original generating parameters of zero. You also notice a skew to the distribution, this skew will be dis-

cussed in more detail later.

Figure 2 contains a line plot of the residuals for three items. The line plot is simply a frequency distribution. The means and standard deviations of the residuals for each item are as expected under the model (0,1). There is a slight skew to each distribution (also seen in Figure 1) but the proportion of residuals falling around each of the three most likely values (-1.2, 0.0, 1.2) is about .33. This relation holds when data fit the model and all measures and calibrations are identical. When the estimates are identical each of the responses is equally likely.

These two figures illustrate that the lower limit on the number of residuals possible on one item is determined by the number of response alternatives. The only way that residuals can distribute in a continuous pattern is when the people and items are distributed in their estimates. Restrictions in the spread of measures or calibrations result in clusters of residuals.

In Study 2 the item calibrations are sampled randomly from a uniform distribution with a range of four logits and a mean of zero. The person parameters are sampled randomly from a normal distribution with a mean of zero and standard deviation of one. This simulation investigates residuals from a testing situation in which the people are centered on the instrument while the range in the sets of parameters is nearly identical.

Figure 3 is the residual-by-measure plot for Study 2. There is what I refer to as a "structural skew" for the distribution of residuals (Z's) and it is smooth in contour as the logits stretch from about -2.5 to 2.5. Unlike Figure 1, there is no apparent clustering of the residuals. This is because the range in measure and calibrations produces a nearly continuous distribution of expected scores, on which the distribution of residuals depends.

The Z max and Z min boundary lines serve as guides for revealing the asymptotic nature of the residuals as the person measures and item calibrations diverge from one another. As the measures increase, large + Z's become impossible. As the measures decrease, large - Z's become impossible. But as the measures become more extreme,

surprising responses produce ever greater residuals. This type of plot can be constructed for items and the pattern is exactly opposite of what we see here. You can even build a 3D plot with measures, calibrations and Z's as the axes. I built a cardboard 3D model and while I found it fascinating I've yet to discover its practical value.

Figure 4 contains the line plots for the easiest, most neutral, and hardest items. Skewness and kurtosis effects are most evident in the two extreme items. The neutral item fits a Gaussian distribution nearly perfectly. It is apparent that residuals should not be assumed to fit a normal distribution unless, perhaps, the item is centered on the people, which is the case for the neutral item. When an item is not centered on the people, surprising failures and successes will lead to a skewed and peaked distribution of residuals.

In general there is a positive linear relation between item calibration and the skewness of the residuals. On very easy items able persons respond mostly as expected, contributing small residuals bunched around zero with an occasional large negative residual. On very hard items less able persons respond mostly as expected, contributing small residuals bunched around zero with an occasional large positive residual.

Also, in general, there is a quadratic relation between item calibration and the kurtosis of the residuals. This relation is due to the tendency of residuals to cluster near zero as item calibrations diverge from the mean person measure. Particularly for extreme items, the residuals cluster and form peaked and skewed distributions.

Continuing with our inspection of the distributional properties of the residuals we turn to Figures 5-7. These figures contain rankit plots of the residuals. Figure 5 is a rankit plot for the hardest item (d=1.85). The Z's are ordered and these observed order statistics are plotted against their expected statistics. Here we see there are too many large positive residuals, clustering on the negative side of zero, and too few large negative residuals. Figure 6 is the rankit plot for the neutral item (d=.09). It is as nearly "normal" as one is likely to see.

Figure 7 is a rankit plot of the easiest item (d=1.53). We note too few positive residuals, a clustering on the positive side of zero, and too many large negative residuals. These probability plots support our rough interpretation based on the line plots. Now are these unusual patterns? Only if we are expecting normally distributed residuals. But since these residuals are from data which do fit the model their patterns constitute the standard of comparison not the normal distributions. Again, residuals, even from data which fit the model, can not be expected to distribute normally.

A line plot of residuals may suggest that residuals are roughly continuous in their distribution. That is not actually the case. The data observed are categorical responses-only the expected responses approach continuity. The extent of continuity in the expected response depends on the variation in the person and item estimates. In the extreme case where all measures=calibrations, there is only one expected response (as we saw earlier). Most data, however, yield expected responses which are close to continuous. The consequence of a categorical observed response - a continuous expected response is an approximately continuous residual distribution that is composed of discrete "layers" of residuals. Each categorical response contributes a layer. The residuals in a line plot, therefore, can be separated according to the number of response possibilities for the item.

Figures 8-10 plot the residuals against the person measures for the same three items we have been considering. The same residuals are also shown back in Figure 4. Consider Figure 8, for the hard item, here we see three rather distinct patterns, or layers of residuals. The upper-most residuals can from "2" responses. The middle layer of residuals came from "1" responses and the lower level residuals came from "0" responses. Now from this we can tell that in Figure 4, the largest +Z can be identified as having come from a "2" response by a mid-range ability person. We also can tell that on this hard item most persons gave the "0" response, as expected, incurring small residuals. Figure 9 shows the pattern for the centered item, which had a nearly normal distribution of residuals. Here each of the responses is repre-

sented about equally. Figure 10 shows the pattern for the easiest item. Most persons did as expected by scoring a "2" while the large - Z's are due to "1" responses.

This type of plot is useful for revealing which responses by what range of abilit estimate led to surprising residuals, and how frequently which responses were being used at various ability levels.

Finally, Study 3 again uses item calibrations sampled from a uniform distribution with a range of four logits and a mean of zero. But, now the person estimates are sampled randomly from a normal distribution with a mean of one keeping, still, a standard deviation of one. This simulation investigates residuals from a testing situation in which the mean measure of the sample is greater than the mean calibration of the instrument while the standard deviation in the sets of estimates is nearly identical.

Figure 11 is the residual-by-measure plot. The structural skew of the residuals is evident but is exaggerated in the positive direction and truncated in the negative direction of the person measure axis. This is because the mean measure of the people is located one logit above the mean calibration of the item.

The general form of the distribution is identical to that for Study 2 in Figure 3 but here a mistake by an able person produces a residual of greater magnitude than the residual for a less able person who scores a surprising success. We would get the opposite pattern, though the same form, if the mean measure was less than the mean item calibration.

Figure 12 contains the line plots for the easiest, most neutral, and hardest items. Again, the skew and kurtosis relations are evident, only more so. The large gap in the distribution for item #2 suggests that the item is operating as if there were only two rather than three response categories. Scores are either "2" leading to a small positive residual or "0" leading to a large negative residual. A "layered" plot like Figures 8-10 would reveal the true situation. The distribution of residuals is most nearly symmetric for the neutral item; it's difficulty is nearest the mean

measure of the sample, which was 1. The hardest item is only slightly harder than the mean measure. Hence, the distribution of residuals contains relatively slight skew and kurtosis effects.

In conclusion, these three idealized similations reveal how the general distribution of the residuals is affected by the spread of the person and item estimates, and the difference between the mean estimates. Other factors influencing the shape of the distribution include the number of persons, items, and response categories.

These studies illustrate that "unusual" structures are the norm, and that these structures can be predicted. Furthermore, they illustrate that Rasch model residuals cannot be assumed to follow a normal distribution, except under very strict circumstances.

The modelled asymmetry of the residuals effects how observed data residuals should be interpreted. For interpreting the fit of observed data to the model, it is not enough to note that an item or person incurred a large residuals. Since asymmetry in the distribution of residuals can occur as a consequence of the model, there must be evidence that the appearance of large residuals is pattern-disrupting and unexpected before model misfit can be claimed. Large residuals may be very informative about an item or person but their appearance does not necessarily mean something is wrong! Thus, any analysis of observed residual variation can only be undertaken by comparing their patterns to those from residuals from data generated to simulate the observed data as closely as possible. An assumed, hypothetical distribution of person and item estimates is an inappropriate background for analyzing observed residuals.

Now, two examples that illustrate some of the practical significance of analyzing observed data residuals in concert with residuals tailored to the testing situation.

The first example discusses an instrument constructed to measure attitudes toward blindness. There are 19 items, 222 persons, and 4 response categories: strongly agree=3, agree=2, disagree=1, strongly disagree=0,(the higher the score, the more positive the attitude.) Three interviewers collected the data from blind

patients who participated in a blind rehabilitation program at the United States Veterans Administration Hines Hospital, Hines, Illinois. The instrument was administered prior to participating, immediately after release, and six months after release. The data are calibrated with the Rating Scale model and, given the person and item estimates, multiple sets of tailored data were generated and analyzed.

Some of the items include the following:

1. A blind person can be a superior piano tuner.
2. Blind people are more honest than sighted people.
3. Blind workers complain less.
4. A blind person develops extra senses.
5. A blind person can raise a normal child.

Since each simulation is one "what if" event, it is, obviously, prudent to replicate simulations. Otherwise, one runs the risk of treating a single simulation as "truth" and then building an analysis around discrepancies between that single case and the observed data. The risk in this strategy is that the single simulation might not resemble the mean pattern of additional replications.

The problem is how many should be done, and do you compare multiple plots to one another? What I do is generate three sets of tailored data, generate my plots, and compare each set separately with the real data. Then see what consensus or differences exist between the tailored data sets. Obviously, there is a degree of subjectivity involved.

Figure 13 plots the residuals from the tailored data against the person measures. The circled area highlights a part of the expected pattern that becomes significant when compared with the same area for Figure 14. In Figure 14 we notice a relatively large number of middle-range attitude patients who have provided surprising disagree responses. Their low scoring responses, given their relatively high attitude measures resulted in large negative residuals. But which items are they, and which patients?

To understand these residuals in clearer detail, we can plot the residuals in item sequence order. Figure 15 plots the tailored residuals and reveals the expected

patterns. Figure 16 reveals the observed residual pattern. We see that most of the large negative residuals come from the first few items.

This pattern suggested that some form of "start-up" effect might be influencing the measurement process. The next step, therefore, was to construct line plots of the residuals broken down by time period and interviewer. Figure 17 contains the expected pattern from the tailored data and Figure 18 contains the pattern for the observed data. What is revealed is a pattern of large negative residuals from surprising disagree responses at Time 1.

The pattern for this item is typical for others of the first few items. True, there aren't many residuals here but we decided to check with the interviewers in order to uncover anything unusual in their techniques or patients. When the interviewers were presented with these patterns they explained that most patients did not respond using the original suggested response categories. Instead they responded "right", "false", "true", "sometimes", etc. Patients without strong convictions did not express their attitudes strongly. The interviewers were then required to interpret those responses. After a few items they usually picked up the patients' pattern and distinguished between middling and extreme responses. But each interviewer handled that situation in an idiosyncratic fashion. This "start-up" effect was a systematic source of measurement error at time period 1. It was partially remedied by introducing a few "warm-up" items.

Since many line plots can be constructed in an analysis one simple way to summarize these line plots is to plot pairs of mean residuals for the items. If only two groups are created, then such a plot will have a negative slope because the means will approximately sum to zero. Otherwise they should be scattered about the origin.

(The means of standardized residuals are not expected to sum to zero because the residuals are not standardized relative to a common error term. According to current terminology these residuals are "internally studentized". The transformation of the estimated residual into a scaled residual is accomplished by dividing by a standard error modelled for each expected response.)

Figures 19 and 20 contain plots of pairs of mean residuals for each item for Interviewer A (a man) and Interviewer B (a woman). Their means are not expected to sum to zero because two other means were also computed (Interviewer C at Time 2, Interviewer B at Time 3). Therefore, the pairs of means may lie in any of the four quadrants. This pair of interviewers was selected because discussions with the interviewers suggested that the responses to some questions by some patients were influenced by having to respond to a woman.

Figure 19 contains the set of tailored residual means. If there is no effect of interviewer gender on patient response, then there should be no pattern when Interviewer A means at Time 1 are plotted against Interviewer B means at Time 1. Such a random pattern is seen if Figure 19.

Figure 20, however, contains three points that stand out from the others. In Quadrant II the discrepancy in scoring "work" has already been addressed in terms of "start-up" effect. The relation between "sex" and "marriage", however, is a new piece of information. (A blind person can offer their spouse satisfactory sex) and (Being blind is an asset to marriage). Interviewer B, the woman, elicited surprising negative responses from some male patients on these two items. Their negative responses led to negative residuals. A similar configuration resulted when her means were plotted against the other man at Time 2. The responses she and the two men interviewers elicited on these two items are different. In particular, these two items were hard for some patients to agree with when she conducted the interview.

Further investigation revealed that most of these men had been interviewed by the women after the patients entered the hospital and that these interviews had been conducted in their private rooms. The results of this analysis and anectodal evidence from rehabilitation staff members (regarding the effect that a young woman and older man walking off to a private room had upon the general population) led to a change in interview locale. The multitude of problems in these data suggested that a more global assessment of misfit might be informative.

Figure 21 contains the first two unrotated principal components extracted from the inter-item correlations for the tailored residuals. Here we are concerned with the undimensionality of the instrument. The location of items should be random and without substantive meaning. This is what we interpret from Figure 21. We could make no sense of the configuration.

Figure 22, however, contains the unrotated principal component solution for the observed residuals. The difference between the first two eigenroots of the tailored and observed residuals and the shape of this principal component solution both indicate that a linear structure between the items still remains in the correlation matrix. In the negative direction of the first component are items which generally concern activities that a blind person might be able to do as well or even better than a sighted person. In the positive direction of the first component are items concerning affective characteristics blind persons might gain as a consequence of their blindness. These items question whether a blind person develops positive affective characteristics to a degree that he would not likely have attained if he were sighted. The presence of these item clusters mean that some patients respond to one group of items differently than the way they respond to one group of items differently than the way they respond to the other.

Items in the negative direction ("activity items") include the following (abbreviated):

1. Can be superior piano tuners
2. Can be good supervisors
3. Can participate in group activities
4. Can be sensitive social workers
5. Can offer spouse satisfactory sex.

Items in the positive direction ("affect items") include the following (abbreviated):

1. Can endure boring tasks more easily
2. Are closer to spouse than sighted
3. Blind workers complain less
4. Can understand feelings better than sighted
5. A blind person is an especially Loyal friend.

This lack of unidimensionality was supported when a separate calibration of these item clusters was performed. These data were separated into two sets, each composed of the items in one cluster. Each set was separately calibrated. The pairs of person

attitude measures were plotted. This was done for the observed and tailored data. If the scale is unidimensional, the pairs should fall along a straight line and be fairly highly correlated. If the scale is not unidimensional, the pairs might form any type of pattern.

Figure 23 shows the tailored pattern. Figure 24 shows the observed pattern. As can be readily seen, an overall estimate of attitude is not an accurate measure for a substantial number of persons taking this intrument. Two separate scores are now reported.

The second example discusses DIAL (Developmental Indicators for the Assessment of Learning), an instrument for the screening of gross motor, fine motor, concepts, and communication skills. The function of DIAL is to identify children in need of follow-up services. Only the communication skills scale is analyzed here. There are 8 items, 814 children, and as many as 7 performance levels. The children range in age from 24 and 72 months. They live in three regions of the United States and are stratified by sex and race. The data are calibrated with the Partial Credit model. Three sets of tailored data were generated and analyzed.

Sample items include the following:

1. Articulation of words
3. Remember number, sequence, sentence
5. Name the action presented
8. Number of words in telling of story.

Figure 25 plots the tailored residuals against the children measures. Figure 26 contains the corresponding plot for the observed residuals. The structural skew is evident in each plot but the tailored pattern contains more large positive residuals and fewer large negative residuals than does the observed pattern. Since the tailored residuals come from responses generated to fit the model, these residuals inform us that occasional surprising successes can be expected under the model. But, these successes are not found in the observed data! This is unusual because we usually expect the real data to have greater variation than the simulation data. Each of the tailored data sets gave a similar result, some surprise was expected, but not found.

Given the nature of the administration, and through discussions with the test developers, we concluded that some administrators let their opinions of some children influence

their scoring objectivity. That is, some administrators did not make a serious effort to test younger, less able children on hard items (alledgedly saving mutual time and frustration). This placed a ceiling on the child's ability, denying potential extra credit if all tasks had been offered. This interpretation is supported by Figures 27 and 28 which show that it was the hardest items which were expected to elicit the surprising successes. Figures 26 and 28 also reveal quite a few high ability children who performed less than expected (in the lower right corner).

One reason for the greater number of surprising failures is revealed by examining Figure 29, a table containing sorted residuals. Most of the large negative residuals in the earlier figures are contained in this table. Item L3 requires remembering skills. A series of tasks are presented and a child receives one point for successfully completing each task. The administrator is supposed to mark off each task completed, not just the highest level task completed. All the children under L3 in this table are bright (indicated in the B column), are from the same region (indicated by first digit in ID field), were administered the instrument by the same person (determined by examining the protocols), did successfully complete nearly every step (determined by examining the protocols), but are credited with a very low score. This occurred because the administator marked just the highest level completed. A score of "1" was then entered on a child's data record (at the time of data entry) because only one mark, not three, was recorded. That type of error was easily corrected.

The other item in Figure 29, L1, uses 15 words to test articulation skills, e.g., mouth, sandwich. This is the easiest item for most children to complete. The identification codes again reveal a communality among these surprisingly low scoring children. These children all come from two areas in the southern region of the United States. It is possible that these children are giving a proper Southern articulation to words but the administrators do not have the skill to notice that or, perhaps, they are aware of the accent emphasis and have chosen to score children on a stricter criterion that they assume is more appropriate. That is, the administrator might decide that Southern articulation is not as correct as some imagined "ideal" standard. Here, the use of local scoring

personnel who are aware of regional accents and who score responses relative to the regional standard might help overcome this interaction.

In addition to these specific but relatively minor problems item L5 ("name action") serves as an interesting example of residuals which "overfit" the model. The Wright W-P fit statistic (t= -6.65) indicated that children performed more consistent than expected on this item. That is, few children scored much more or less than expected, given their ability level. The more able kids were, generally, the older kids, they had greater experience. This type of inequity in experience is frequently encountered in "high discrimination" items. Figure 30 (tailored residuals), and Figure 31 (observed) reveal what an unexpectedly consistent performance means in terms of a residual pattern. Now the important feature, here, is the observation that the standard deviation of the observed residuals is less than that for the data tailored to fit the model. This narrow, constricted pattern for the observed residuals is characteristic of items with relatively large negative fit statistics. And, such constricted residual variation leads to high discrimination indices. The response patterns leading to these residuals are more consistent than expected and are flagged by both negative fit statistics and high discrimination values because the residual variation is less than that expected under the model.

In conclusion, the preceding discussion illustrates some of the practical utility of analyzing observed residual variation relative to tailored, or expected, residual variation. The measurement error uncovered in the residual analyses was not noticeable in the examination of person and item estimates, nor the person and item fit statistics. The tailored residuals, coming from data generated to fit the model-given the original estimates, provided a specific frame of reference within which the observed variation would be understood. The use of hypothetical distributions to generate data would not have provided a relevant background for either set of data.

Such an analysis of observed residual variation is expedited when a general systematic strategy is followed. One strategy found useful in our exploratory research entails a hierarchical sifting through of the data. In general, course techniques are employed

first, and person effects are studied before item effects. Those results direct the second level of analysis, and so on, until detailed analyses finish checking all leads suggestive of contributing measurement error. A detailed discussion of a general systematic strategy (including possible patterns uncovered, their meaning and additional steps which might be taken) may be found in Ludlow (1983).

My appreciation is extended to those members of the ETS community who participated in the seminar and offered valuable feedback and interesting suggestions for modifications to my plots and the creation of others.

References

Ludlow, L.H. The analysis of Rasch model residuals. Unpublished doctoral dissertation, University of Chicago, 1983.

On the Simulation and Analysis of Measurement Model Residuals

A presentation prepared for Educational Testing Service

by

Larry H. Ludlow, Ph.D.
Boston College
Jan. 11, 1984

The graphics included in this document were prepared by the software package XPLOR: Expected Plots of Residuals by L.H. Ludlow. The text for this presentation may be found in: The Analysis of Rasch Model Residuals, doctoral dissertation, University of Chicago, 1983, L.H. Ludlow.

No figure is to be reproduced without the prior consent of the author, thank you.

2

General Rasch Model:

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})}$$

$x = 0, 1, \ldots, m$
$M = m+1$ responses

Rating Scale Model:

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} [\beta_n - (\delta_i + \tau_j)]}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} [\beta_n - (\delta_i + \tau_j)]}$$

where:
$x = 0, 1, \ldots m$
$\tau_0 \equiv 0$
$\exp \sum_{j=0}^{0} [\beta_n - (\delta_i + \tau_j)] = 1$

Partial Credit Model:

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})}$$

where:
$x = 0, 1, \ldots, m_i$
$\delta_{i0} \equiv 0$
$\sum_{j=0}^{0} (\beta_n - \delta_{ij}) = 0$
$\exp \sum_{j=0}^{0} (\beta_n - \delta_{ij}) = 1$

Expected value of x_{ni}:

$$E_{ni} = \sum_{k=0}^{m} k \pi_{nik} \qquad k = 0, 1, \ldots, m$$

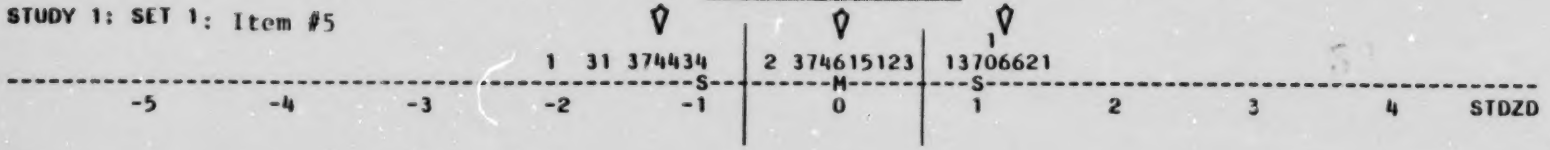
Variance of x_{ni}: $W_{ni} = \sum_{k=0}^{m} (k - E_{ni})^2 \pi_{nik}$

Standardized Residual: $Z_{ni} = \frac{x_{ni} - E_{ni}}{W_{ni}^{1/2}}$

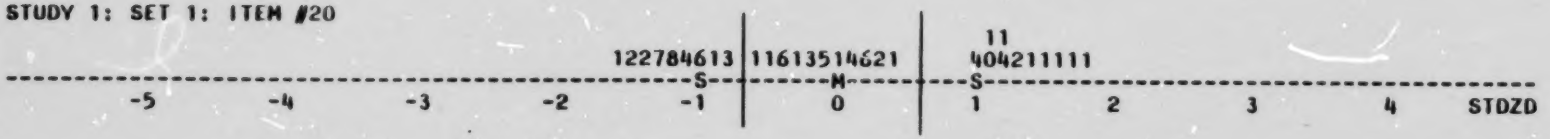
From: Wright, B.D. & Masters, G.N. Rating Scale Analysis. Chicago: MESA Press, 1982.

Three clusters

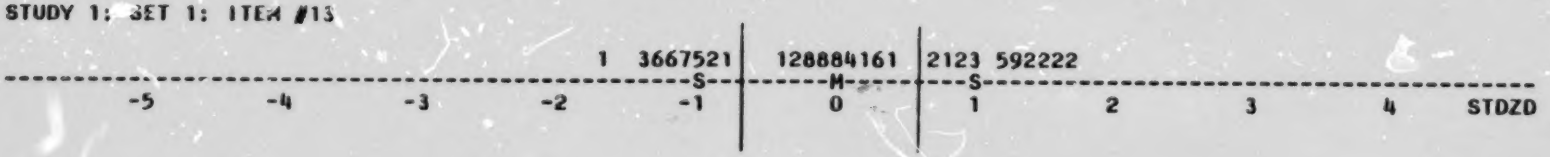
STUDY 1: SET 1: Item #5



STUDY 1: SET 1: ITEM #20



STUDY 1: SET 1: ITEM #13



Summary of statistics

	ITEM 5	20	13
(Calibration)	-0.11	-0.04	0.02
Mean	0.00	0.00	0.00
Standard deviation	1.03	1.01	0.96
Skew	-0.24	0.05	0.12
Kurtosis	-1.40	-1.50	-1.30

Figure 2.--Line plots of three items from Study 1

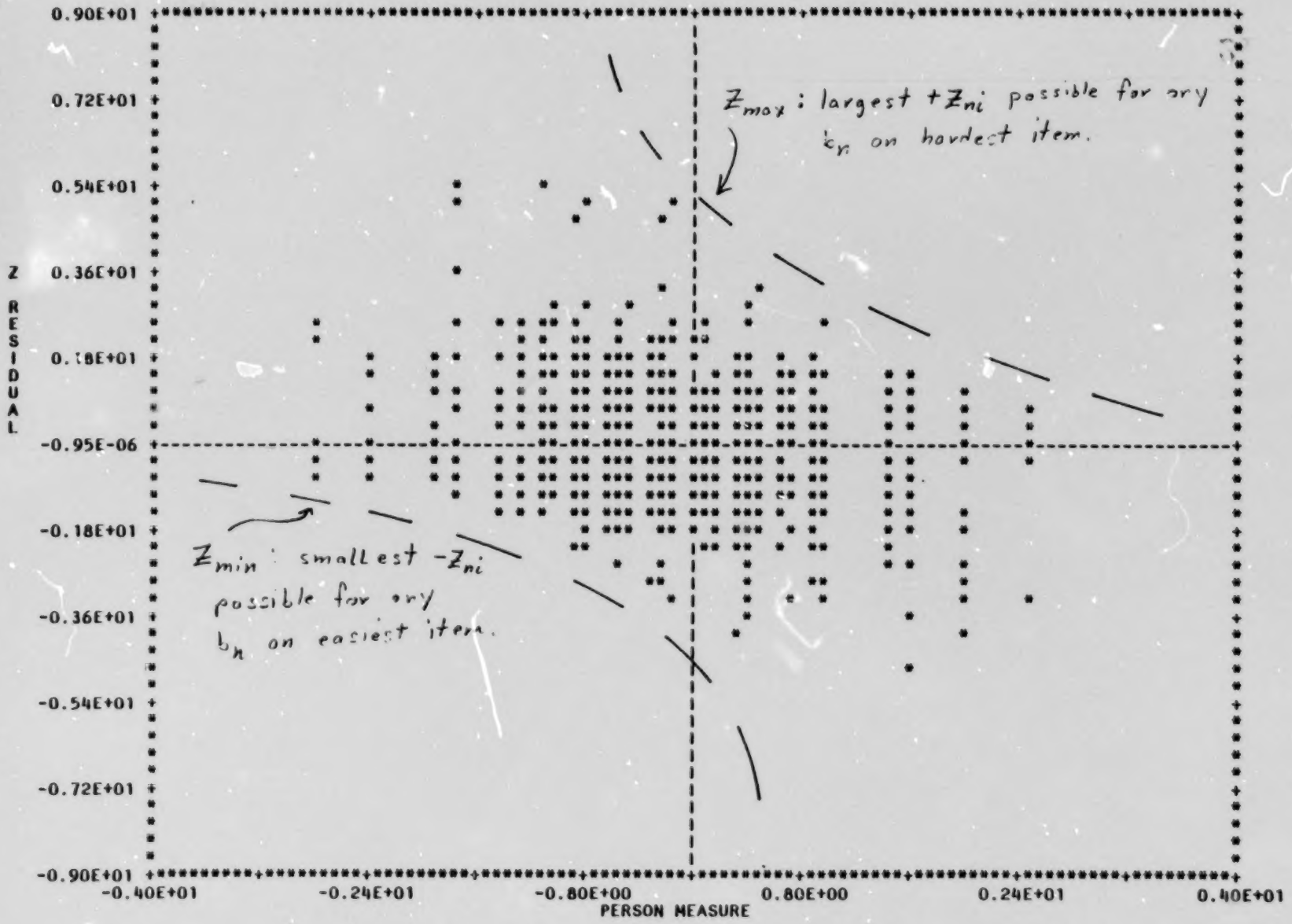
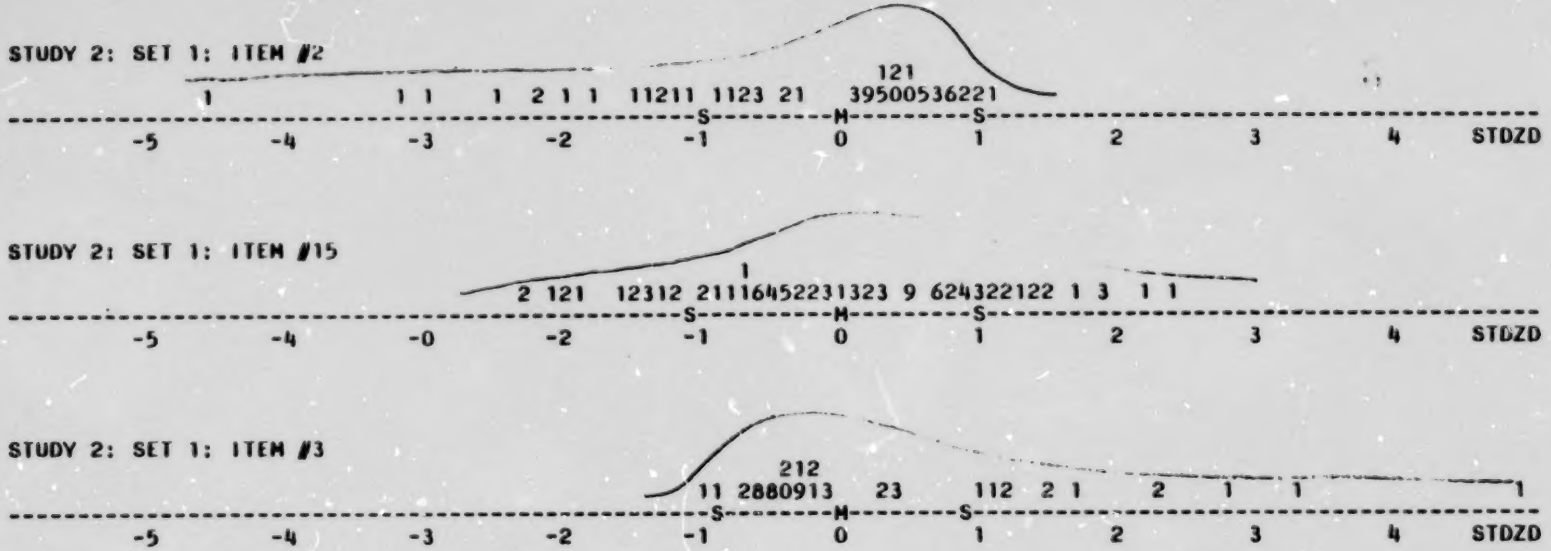


Figure 3 --Residuals versus measures for Study 2



Summary of statistics

	ITEM		
	2	15	3
(Calibration)	-1.53	0.09	1.85
Mean	-0.01	-0.03	-0.04
Standard deviation	0.99	1.05	0.90
Skew	-2.17	0.02	3.15
Kurtosis	5.00	-0.50	10.60

Figure 4--Line plots of three items from Study 2

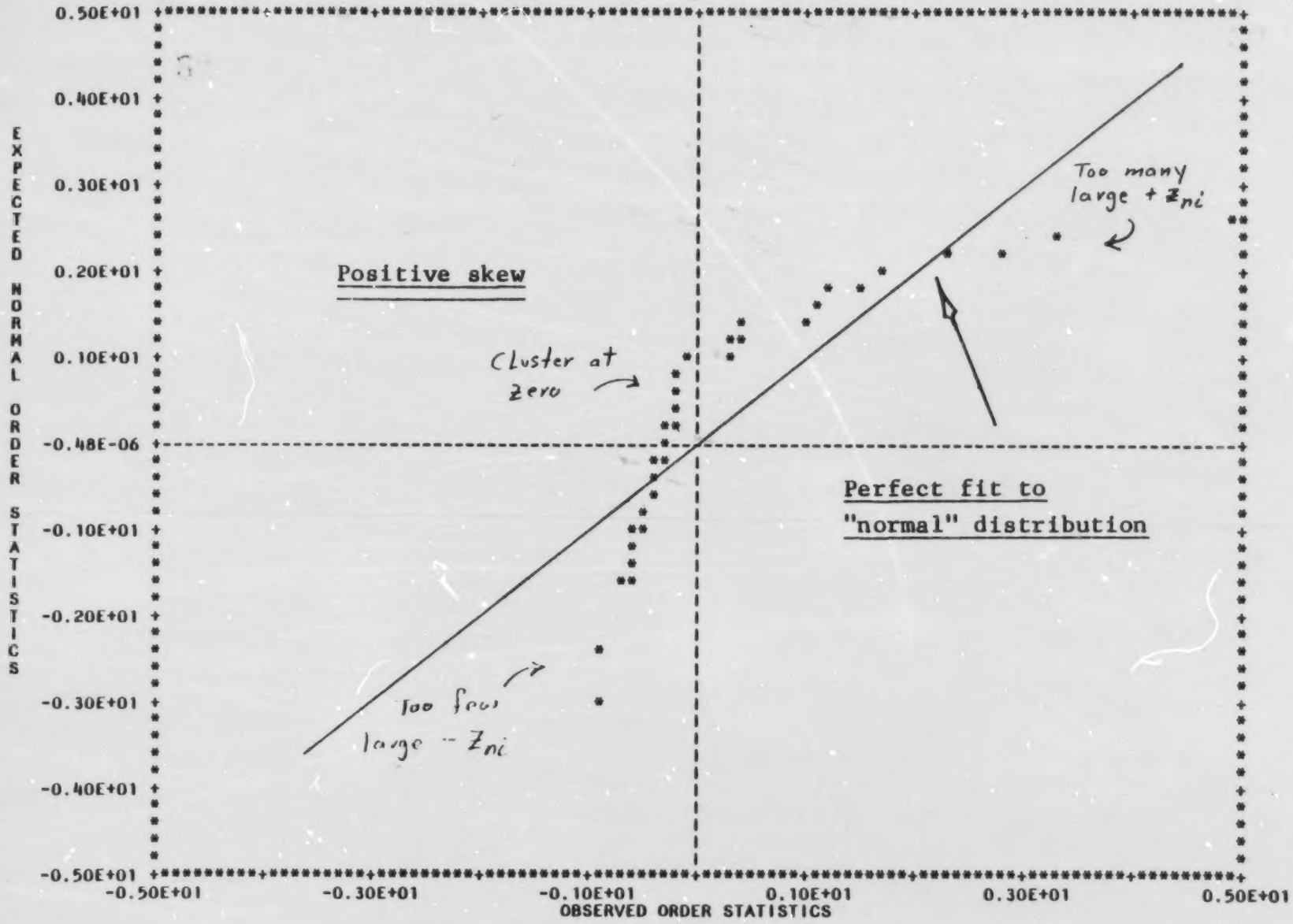


Figure 5 --Rankit plot for hard item from Study 2 (d=1.85)

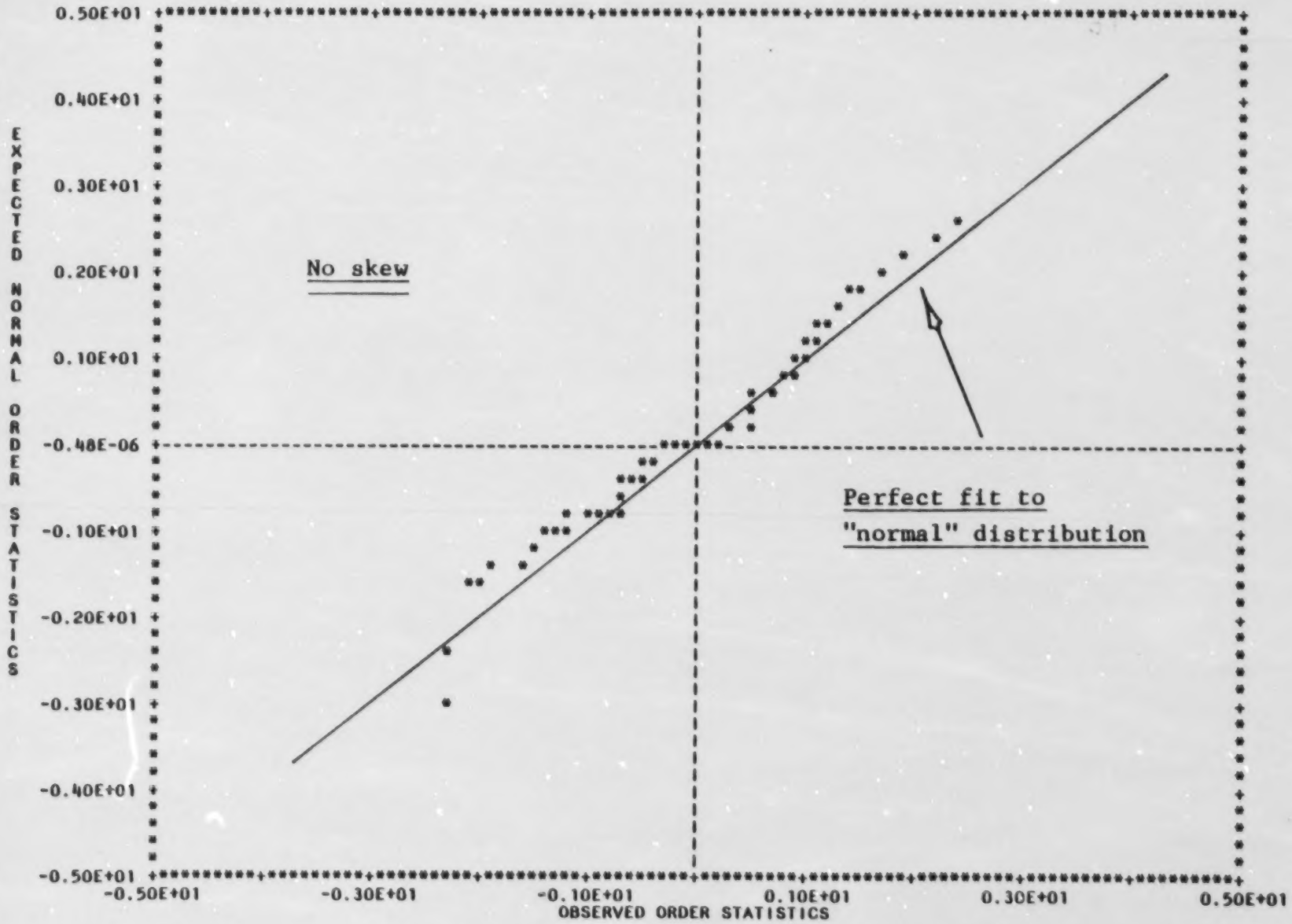


Figure 6 --Rankit plot for neutral item from Study 2 (d=.09)

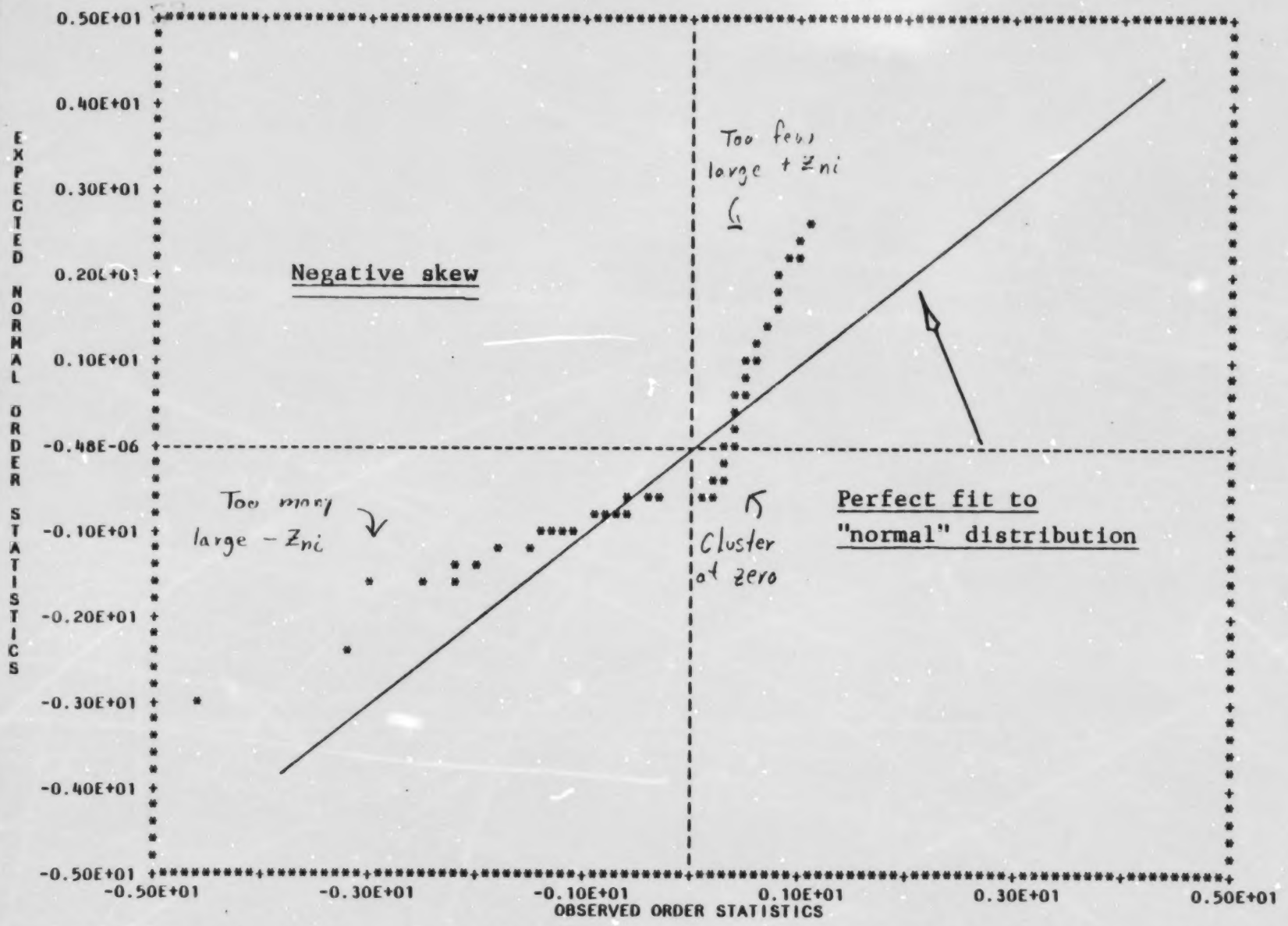


Figure 7 --Rankit plot for easy item from Study 2 (d=-1.53)

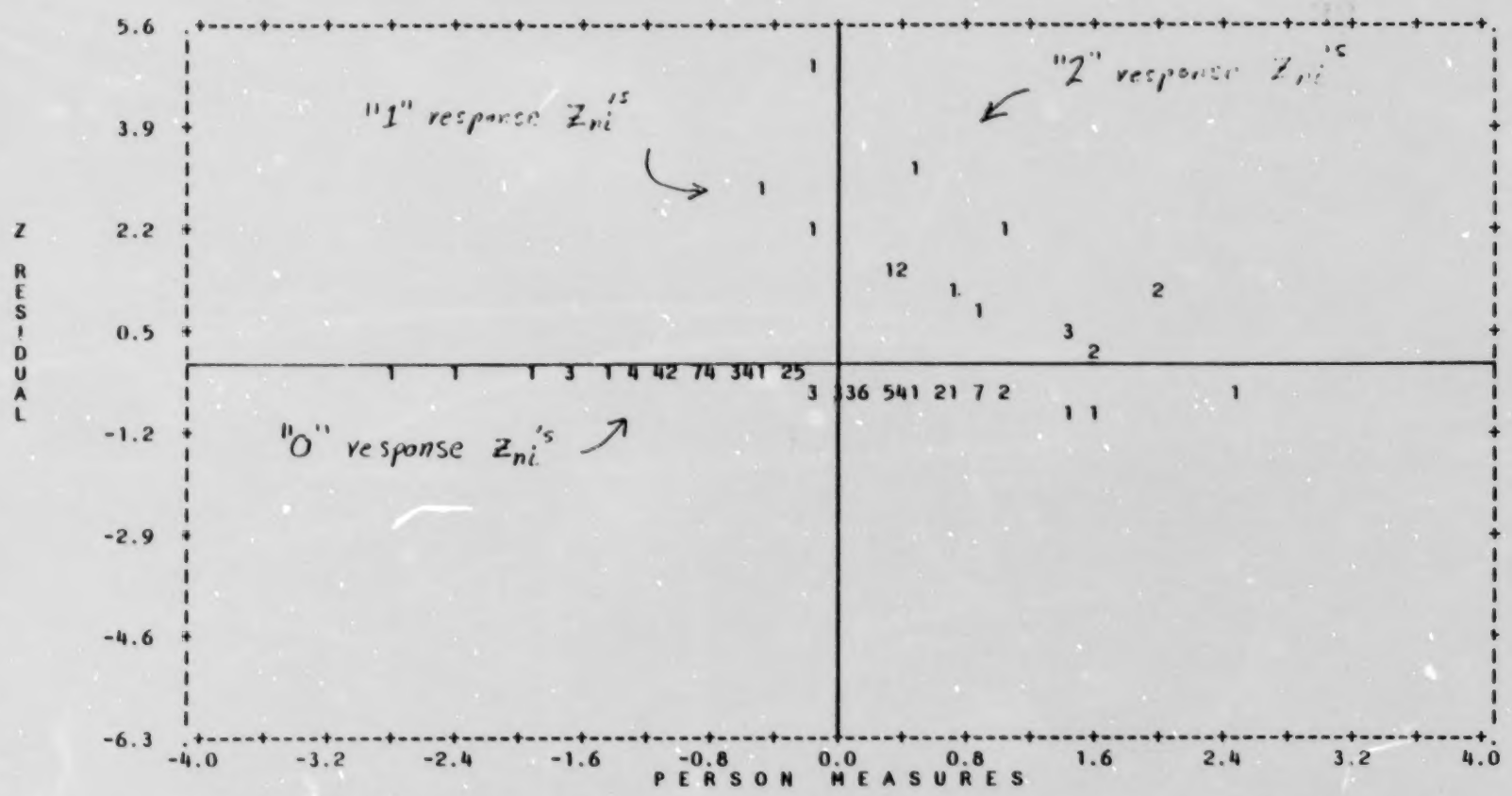


Figure 8 --Standardized residuals versus person measures for a hard item (d=1.85)

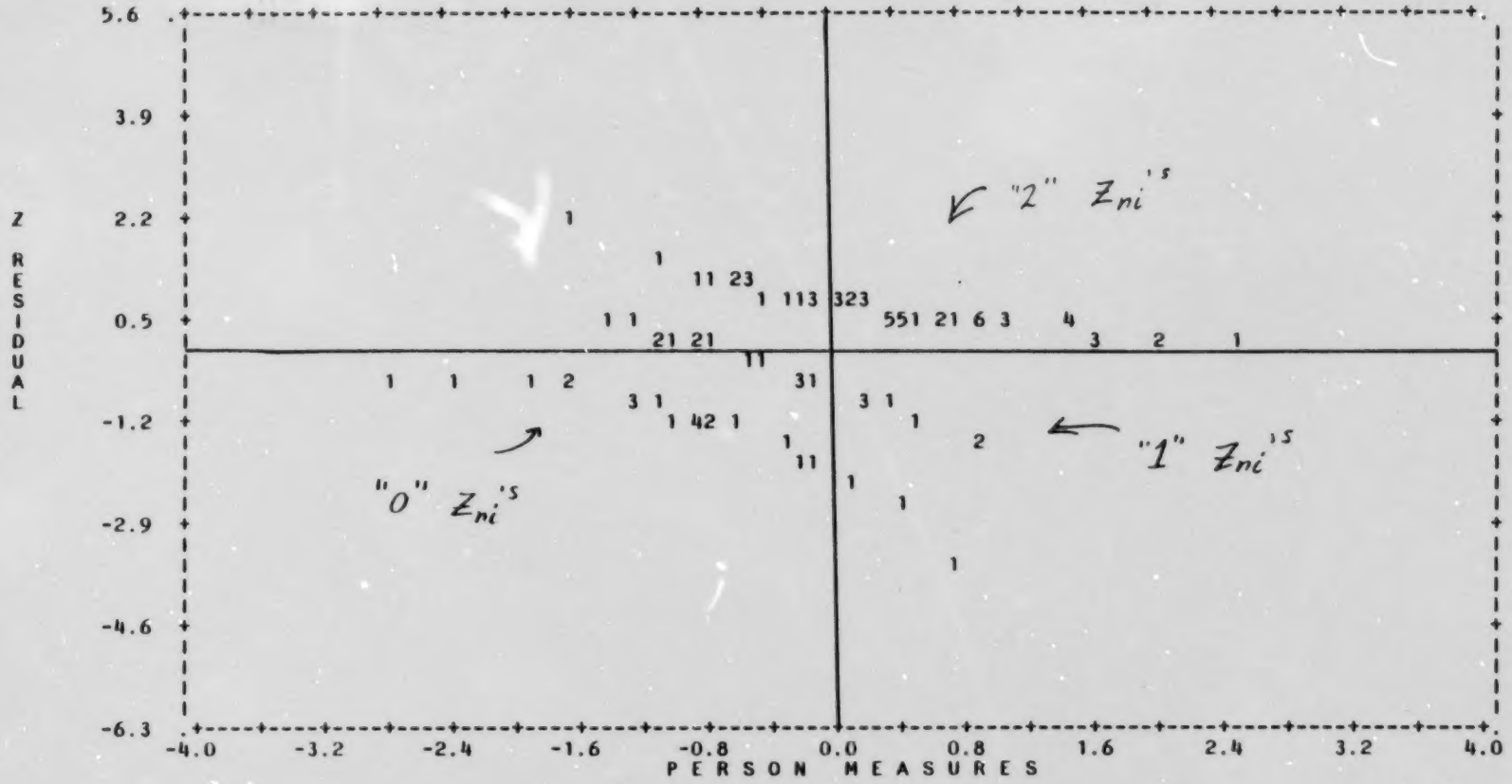


Figure 9 --Standardized residuals versus person measures for a centered item

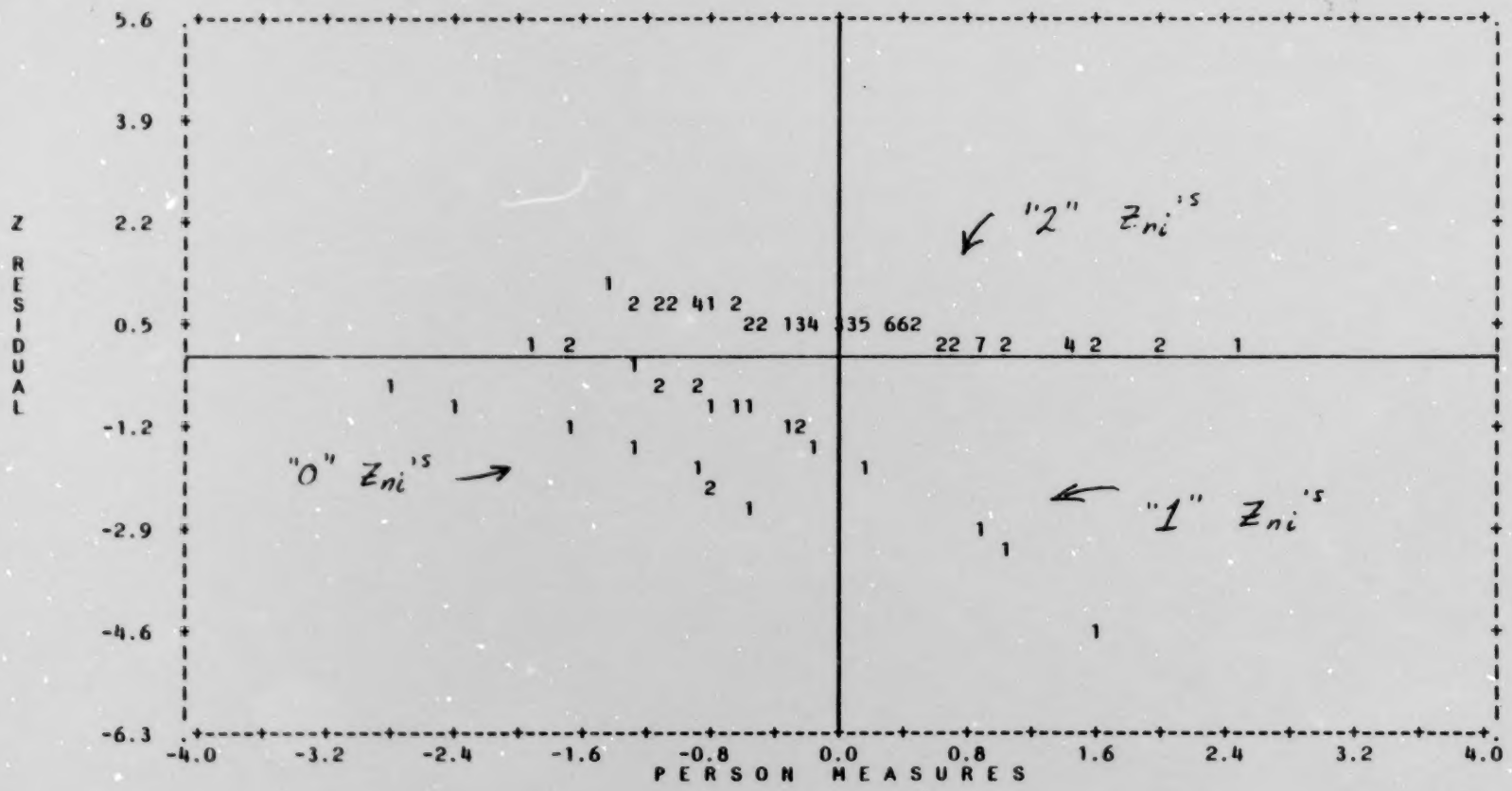


Figure 10--Standardized residuals versus person measures on an easy item (d=-1.53)

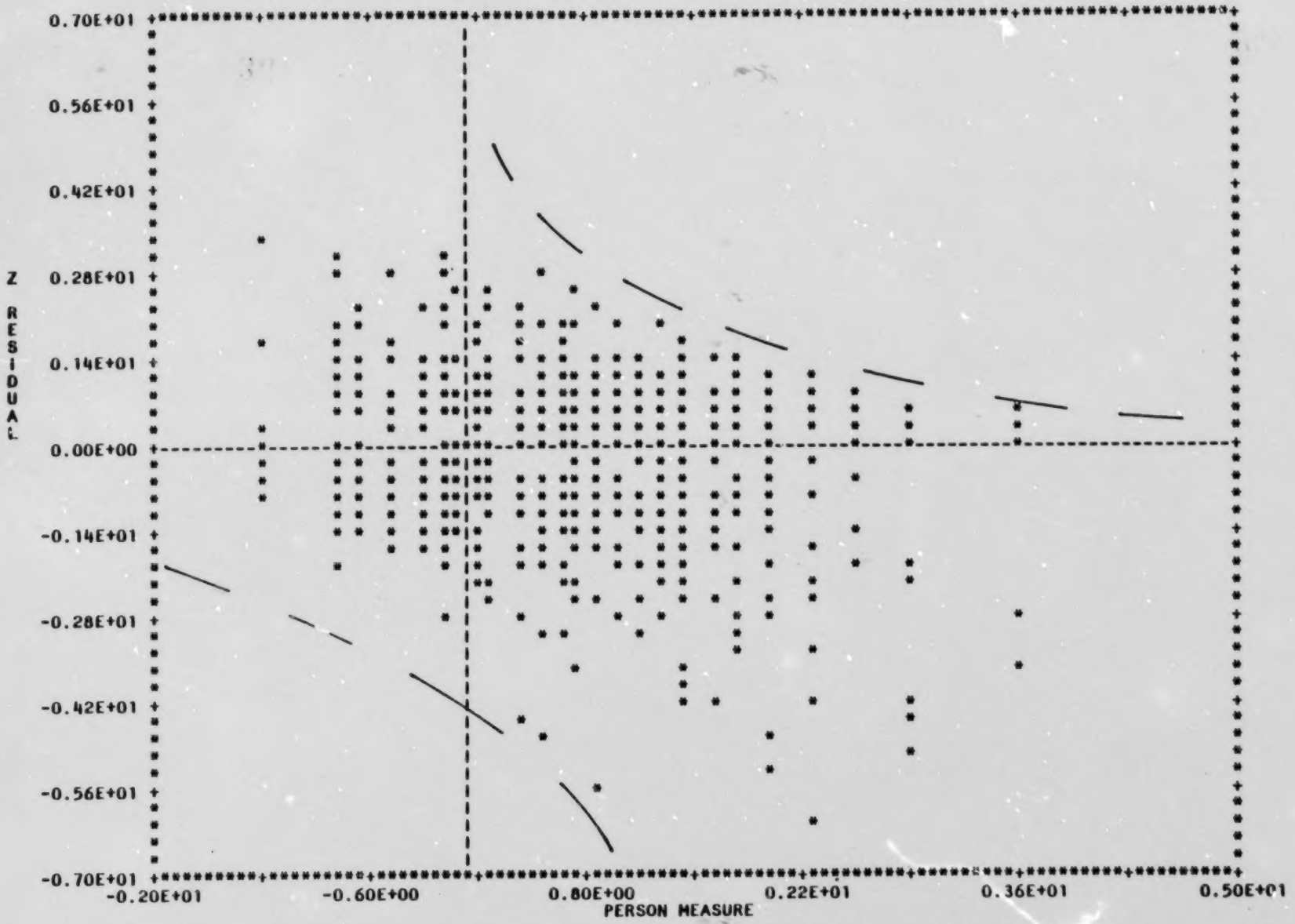
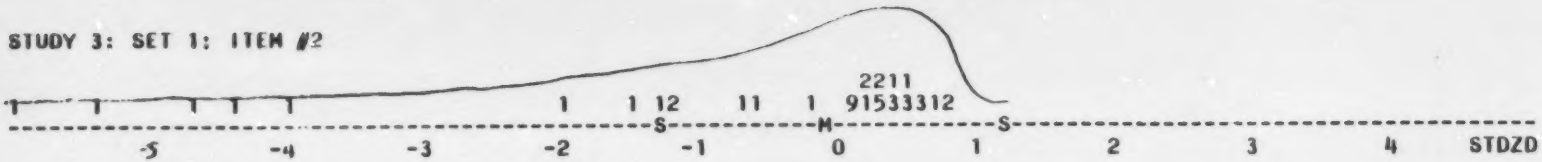
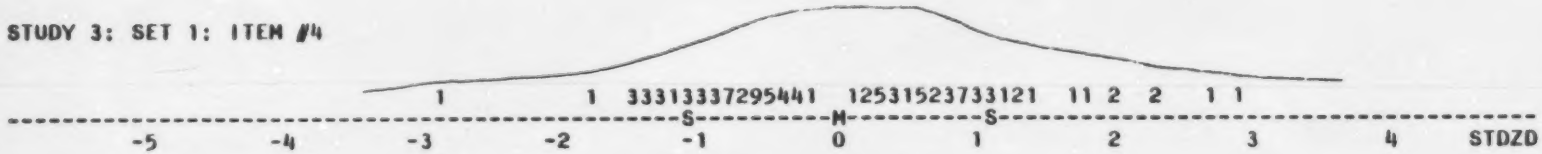
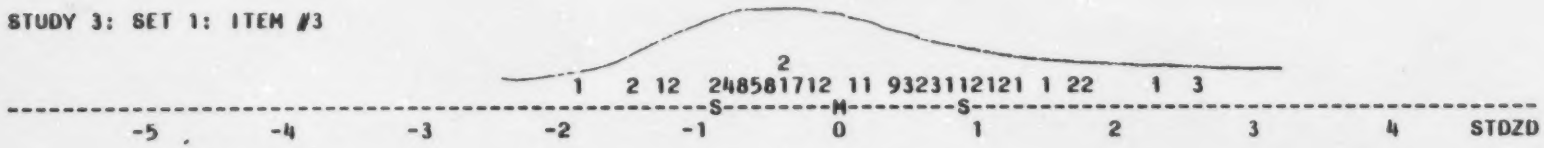


Figure 11--Residuals against measures for Study 3







Summary of statistics

	ITEM 2	ITEM 4	ITEM 3
(Calibration)	-1.45	1.06	1.80
Mean	-0.05	0.01	-0.01
Standard deviation	1.22	1.08	0.91
Skew	-3.27	0.31	1.05
Kurtosis	11.90	-0.20	0.90

Figure 12 --Line plots of three items for Study 3

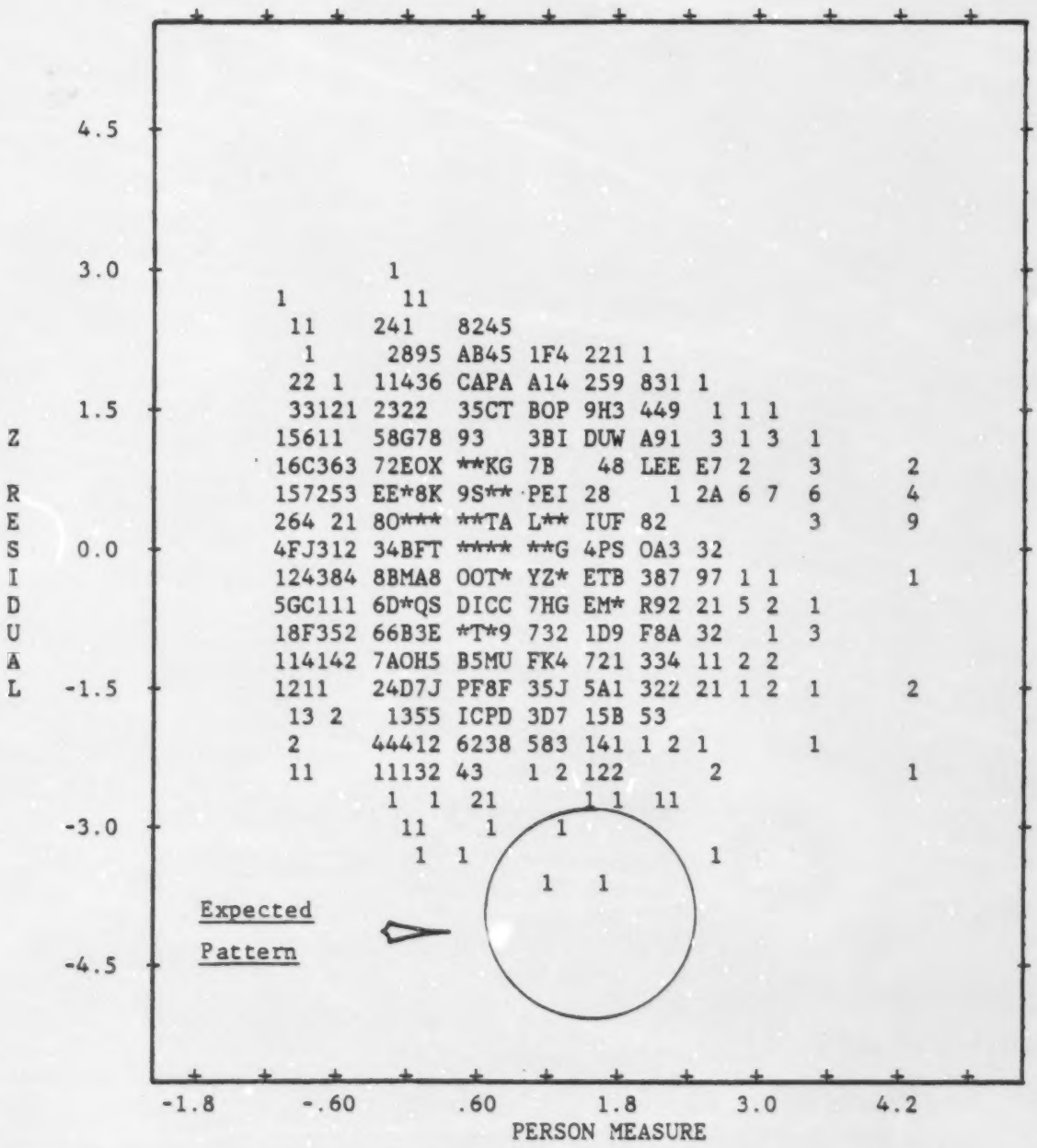


Figure 13--Residuals versus person measures:
Tailored data-first set

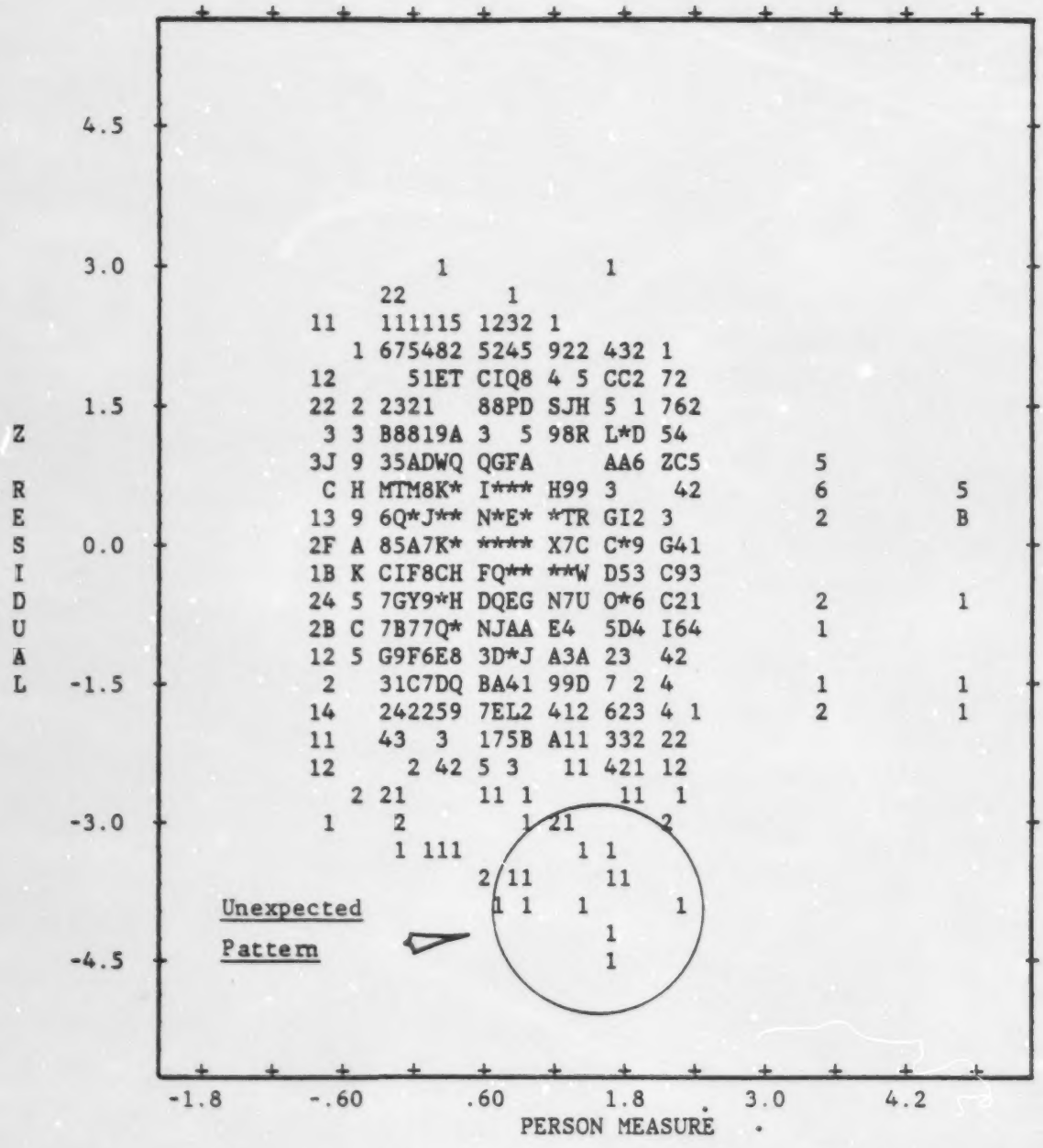


Figure 14--Residuals versus person measures:
Observed Data

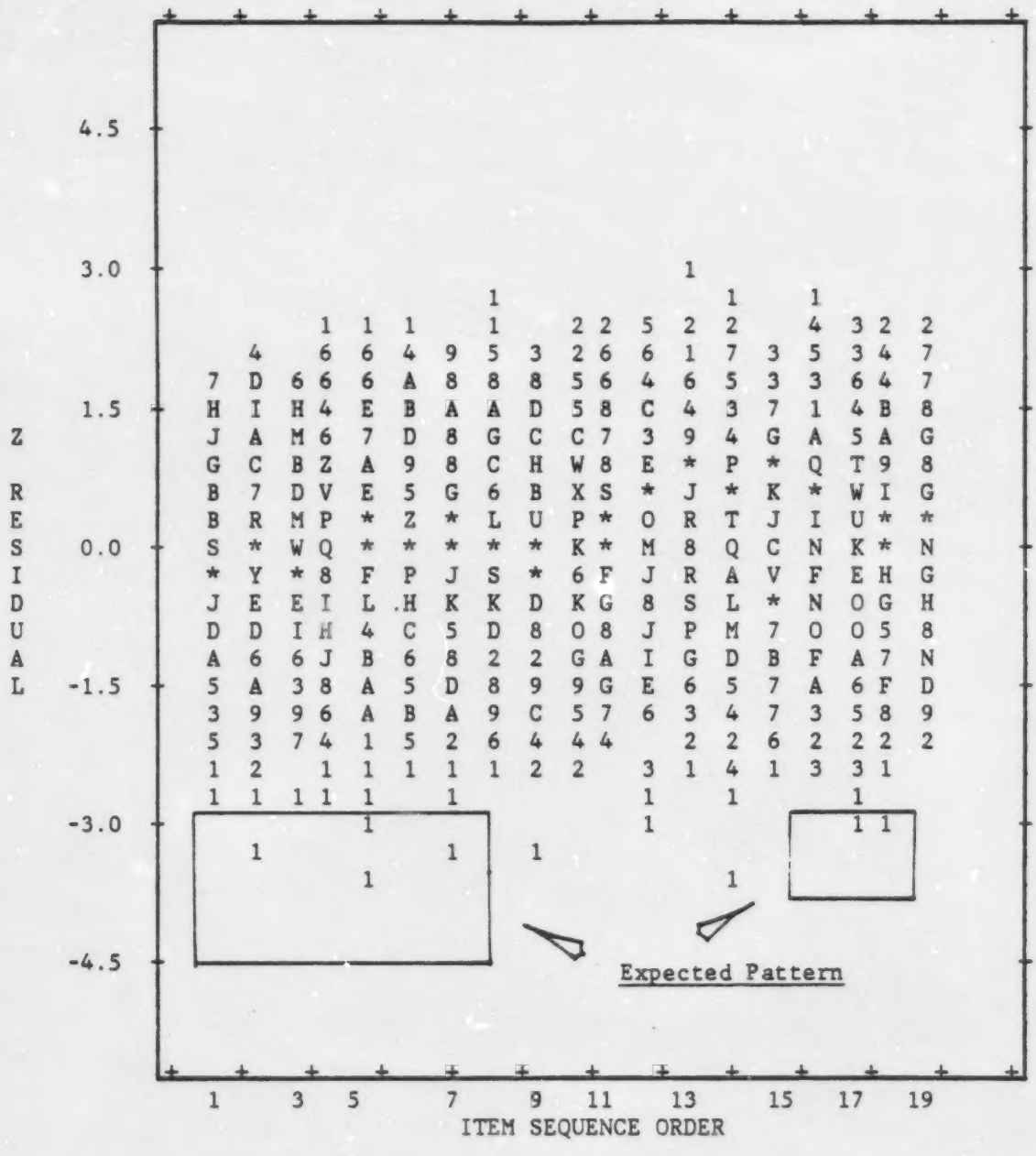


Figure 15 --Residuals versus item sequence:
Tailored data-first set

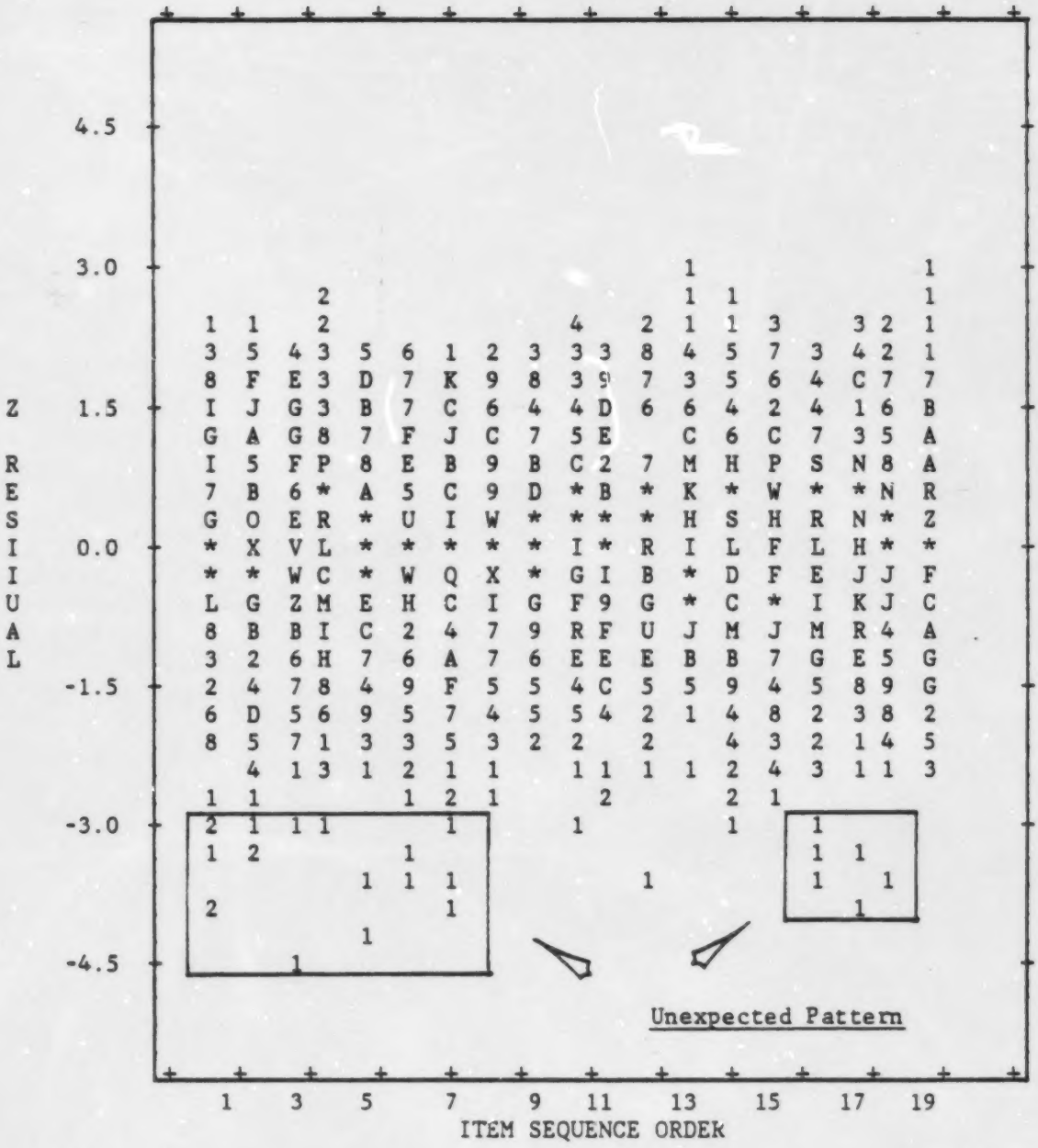


Figure 16 --Residuals versus item sequence: Observed Data

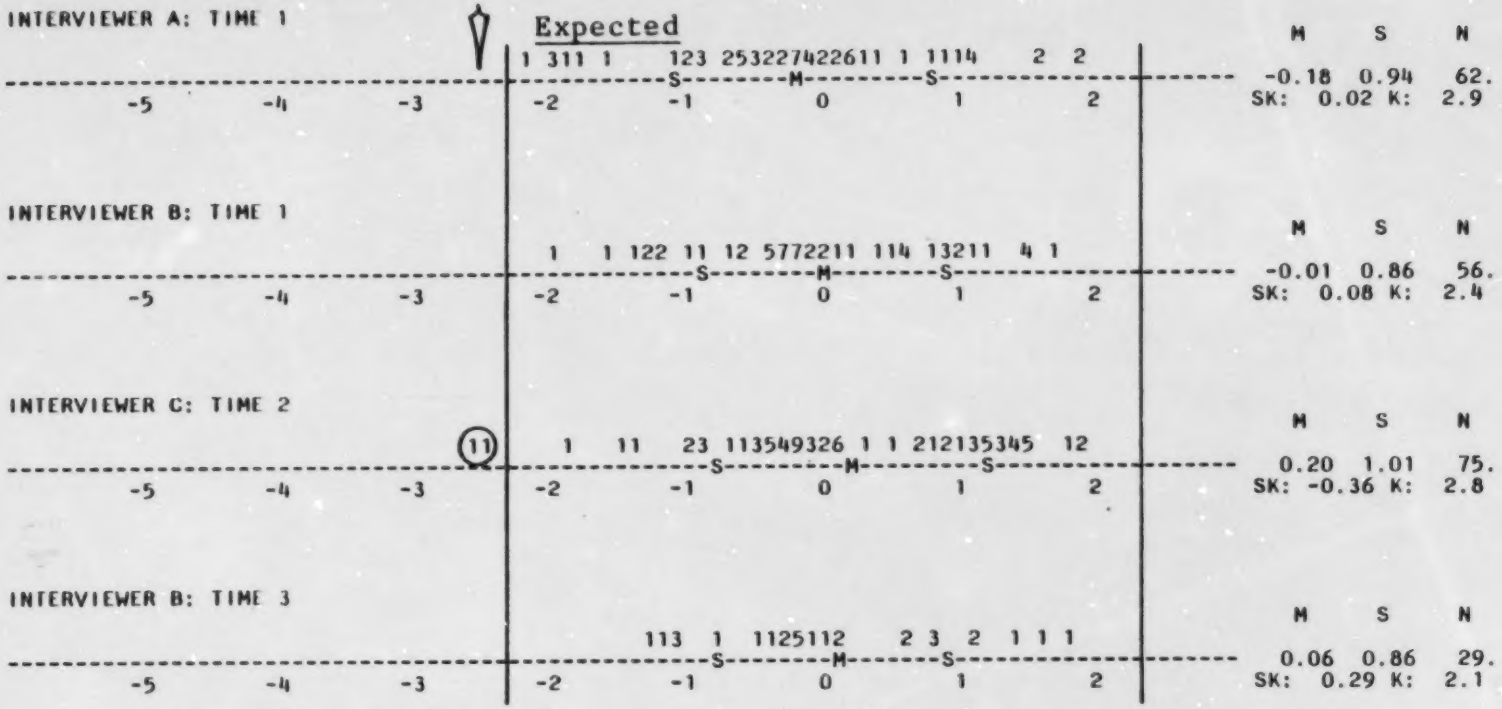


Figure 17--Residuals on I128, "work", by interviewer and time period: Tailored data-first set

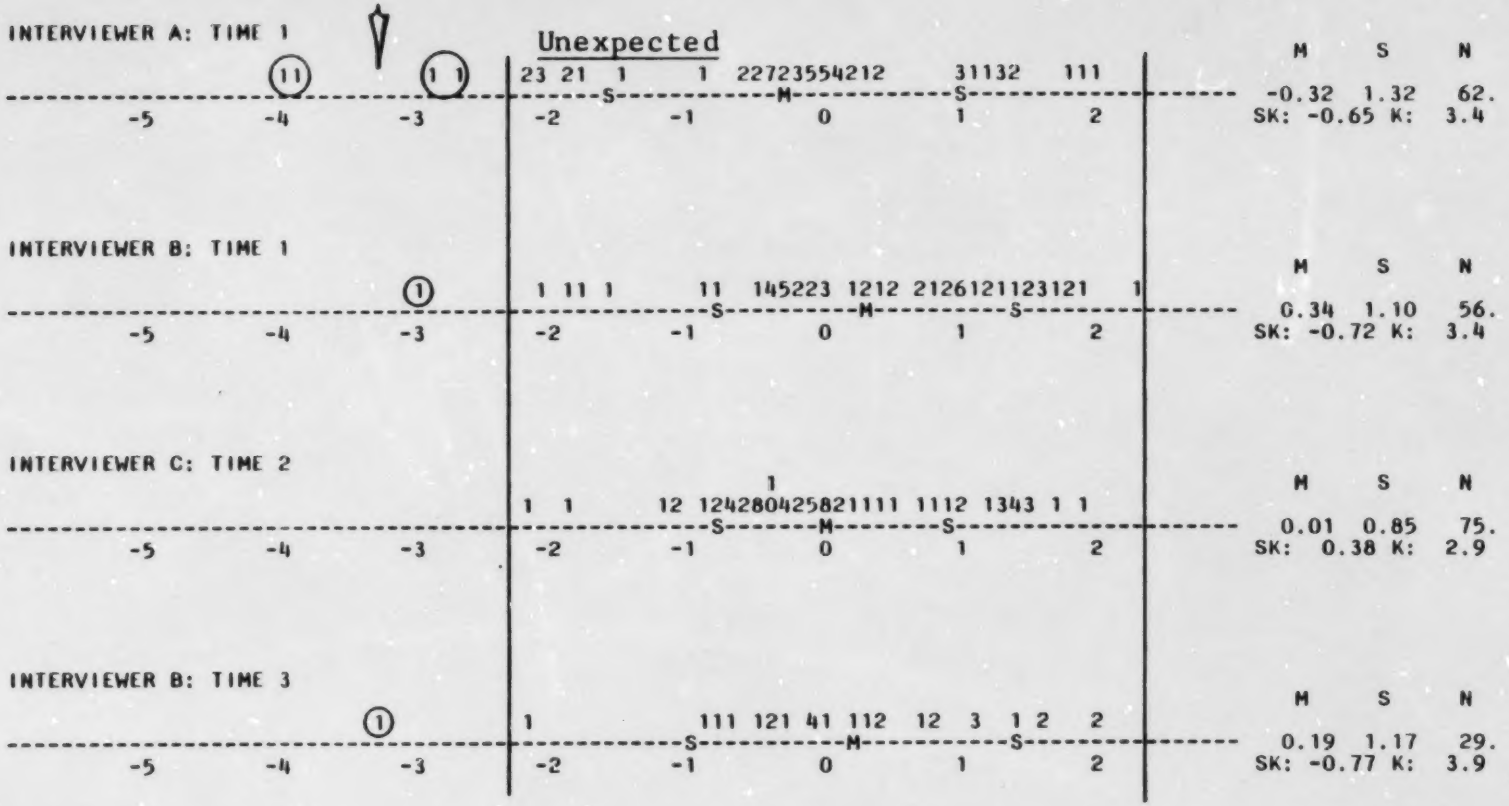


Figure 18--Residuals on I128, "work", by interviewer and time period: Observed Data

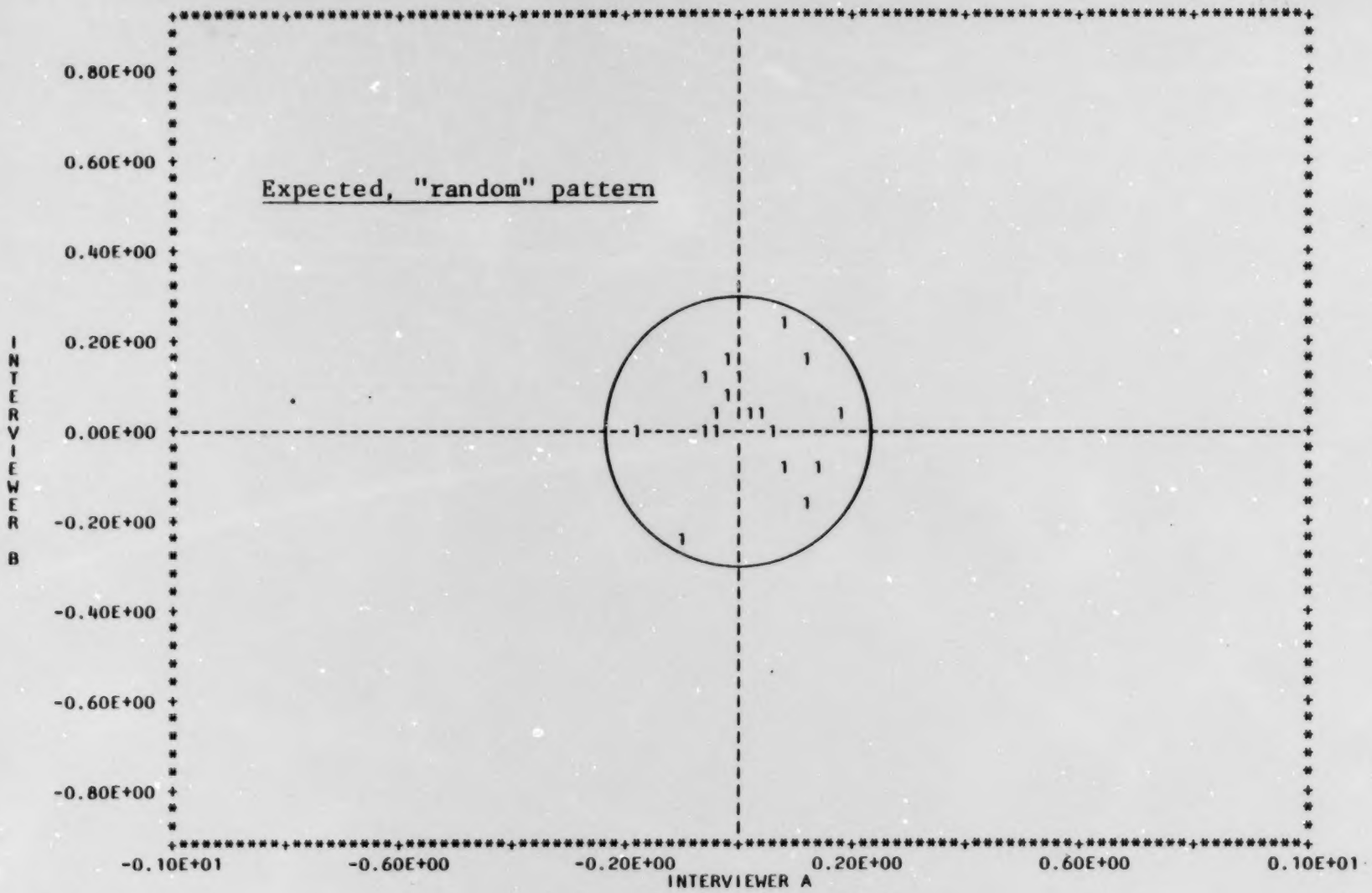


Figure 19--Plot of mean residual on each item for interviewer A and B at Time 1: Tailored data-first set

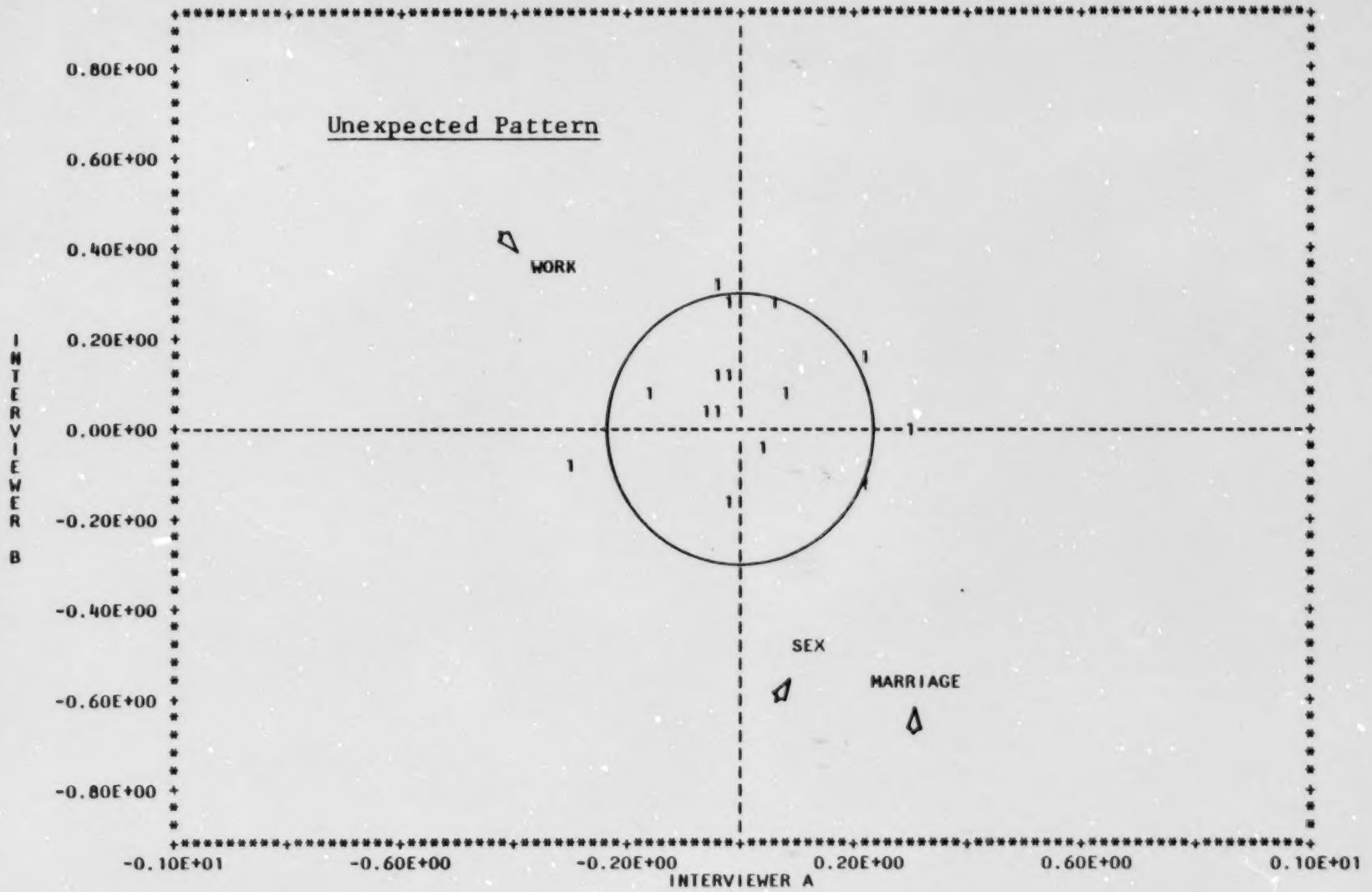


Figure 20--Plot of mean residual on each item for Interviewer A and B at Time 1: Observed Data

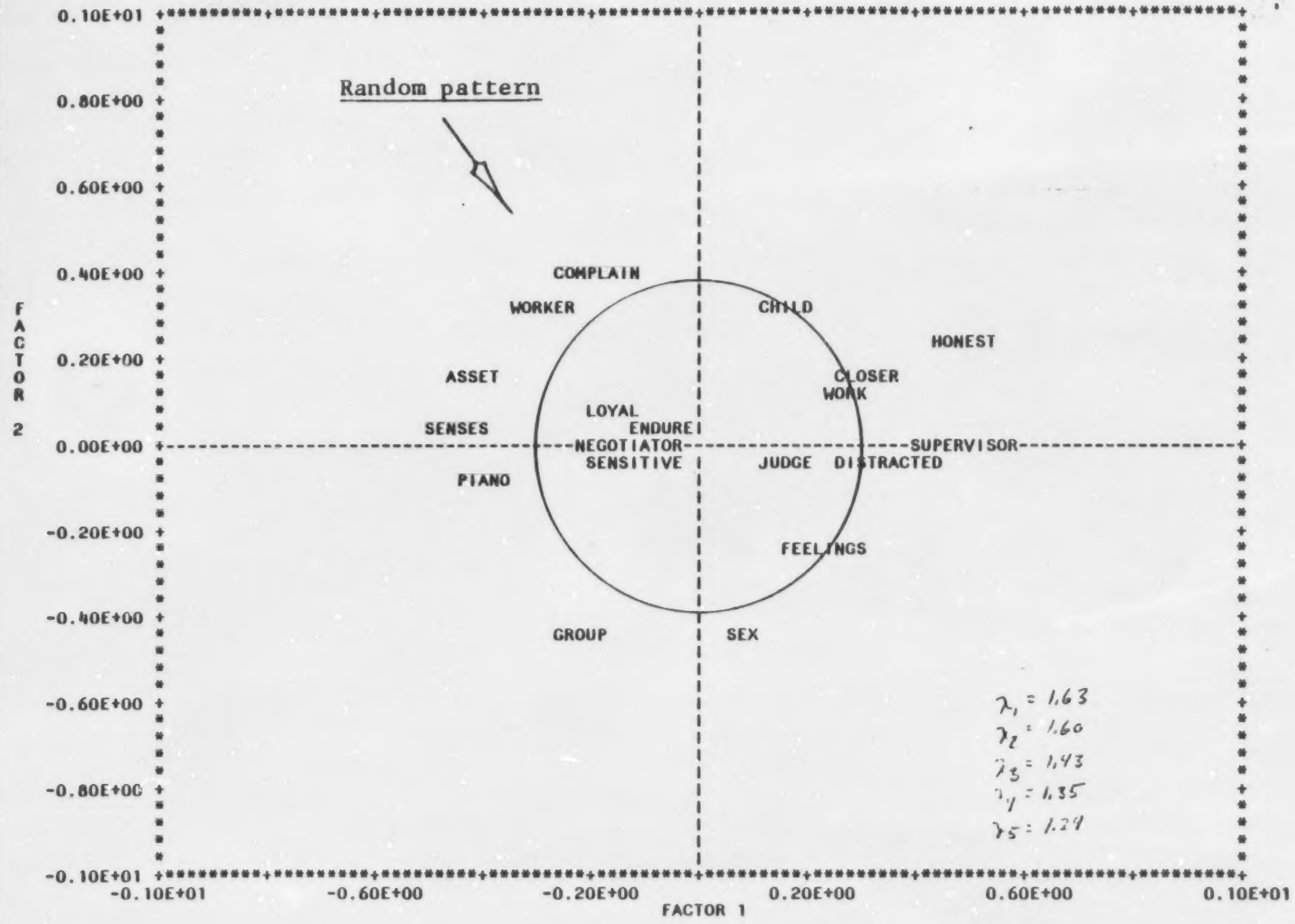


Figure 21--Principal components for "Blind" simulated data-first set

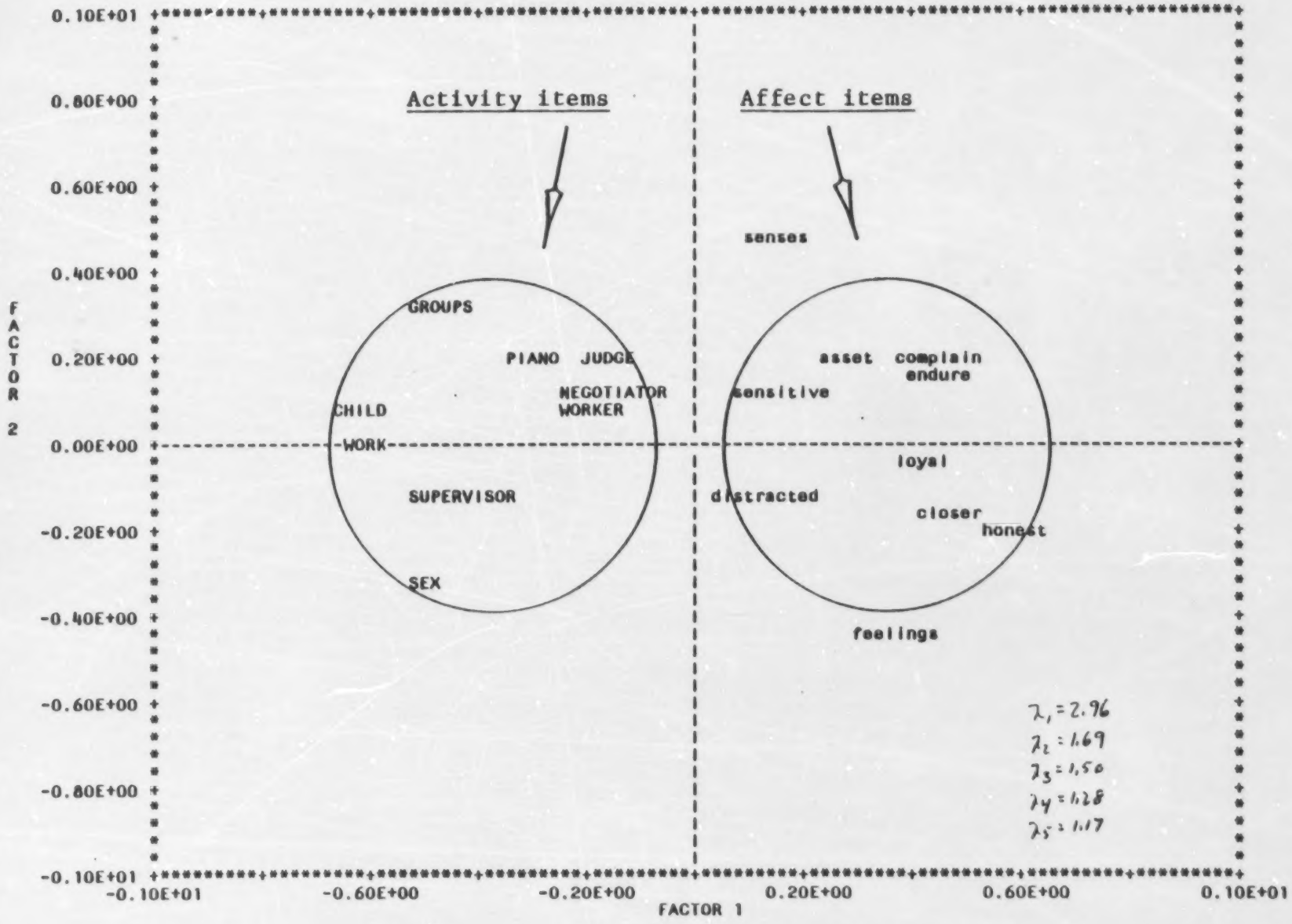


Figure 22--Principal components for "Blind" observed data

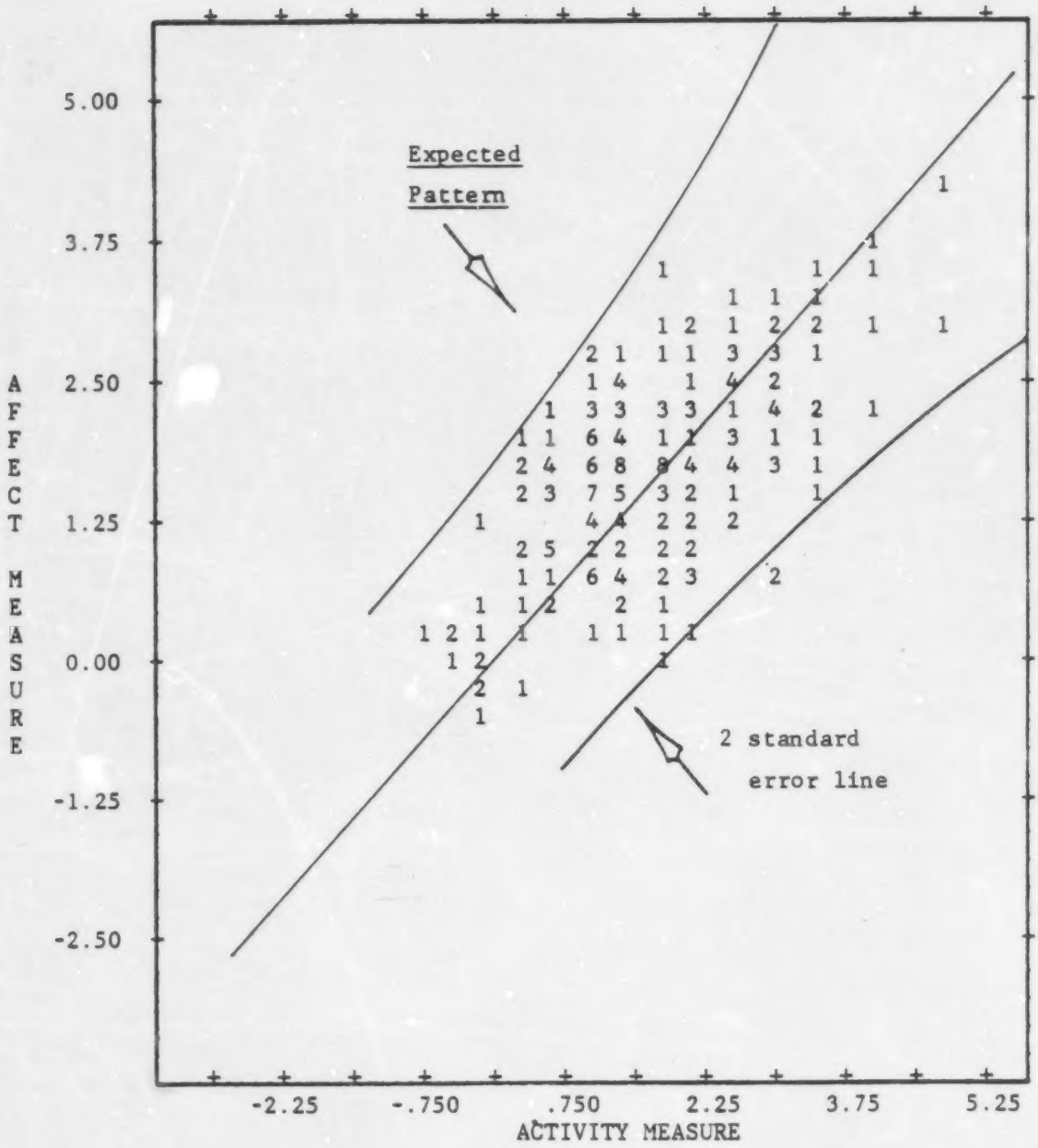


Figure 23 --Affect person measures versus activity person measures: Tailored data-first set (r= .62)

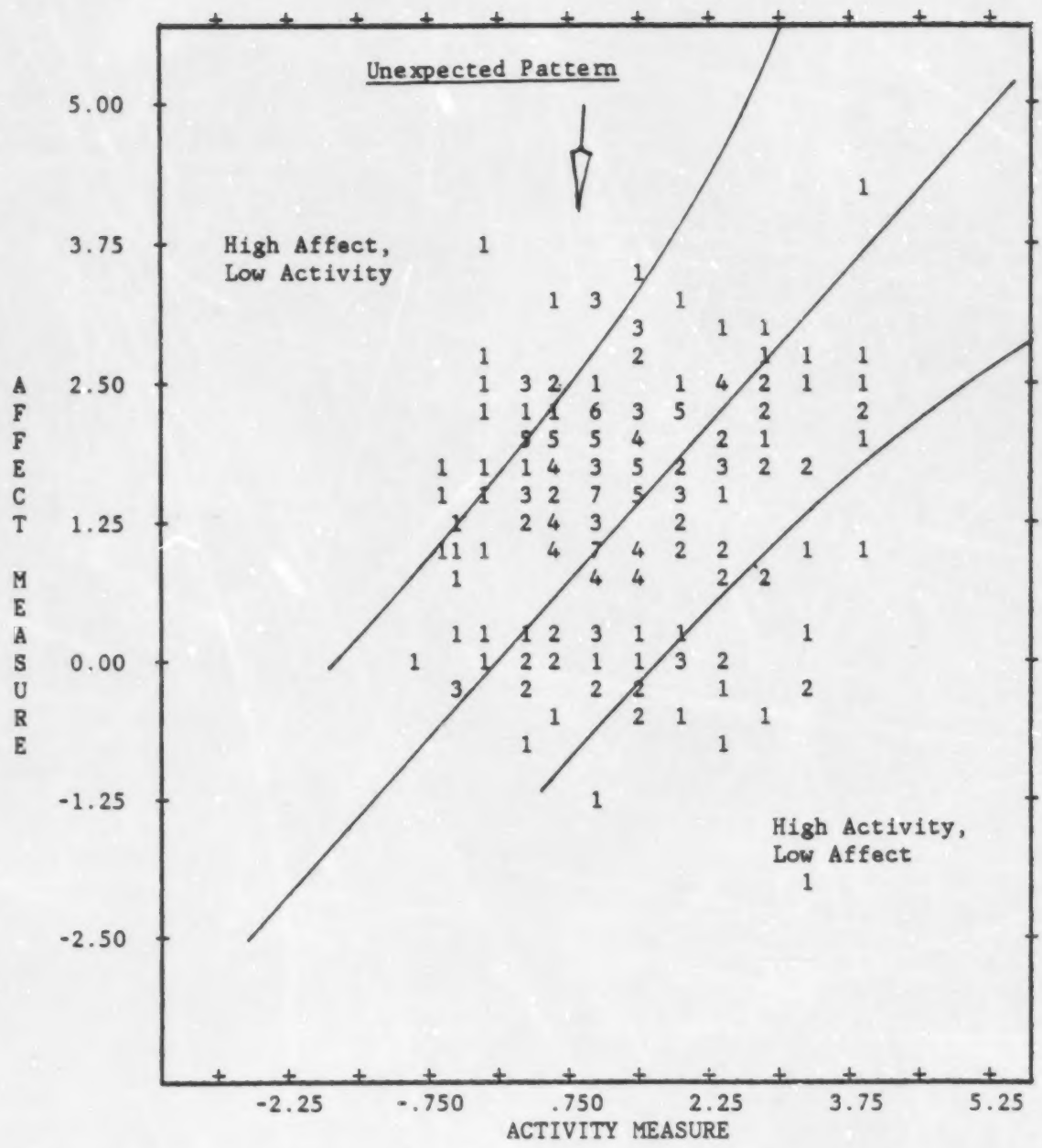


Figure 24--Affect person measures against activity person measures: Observed data (r= .12)

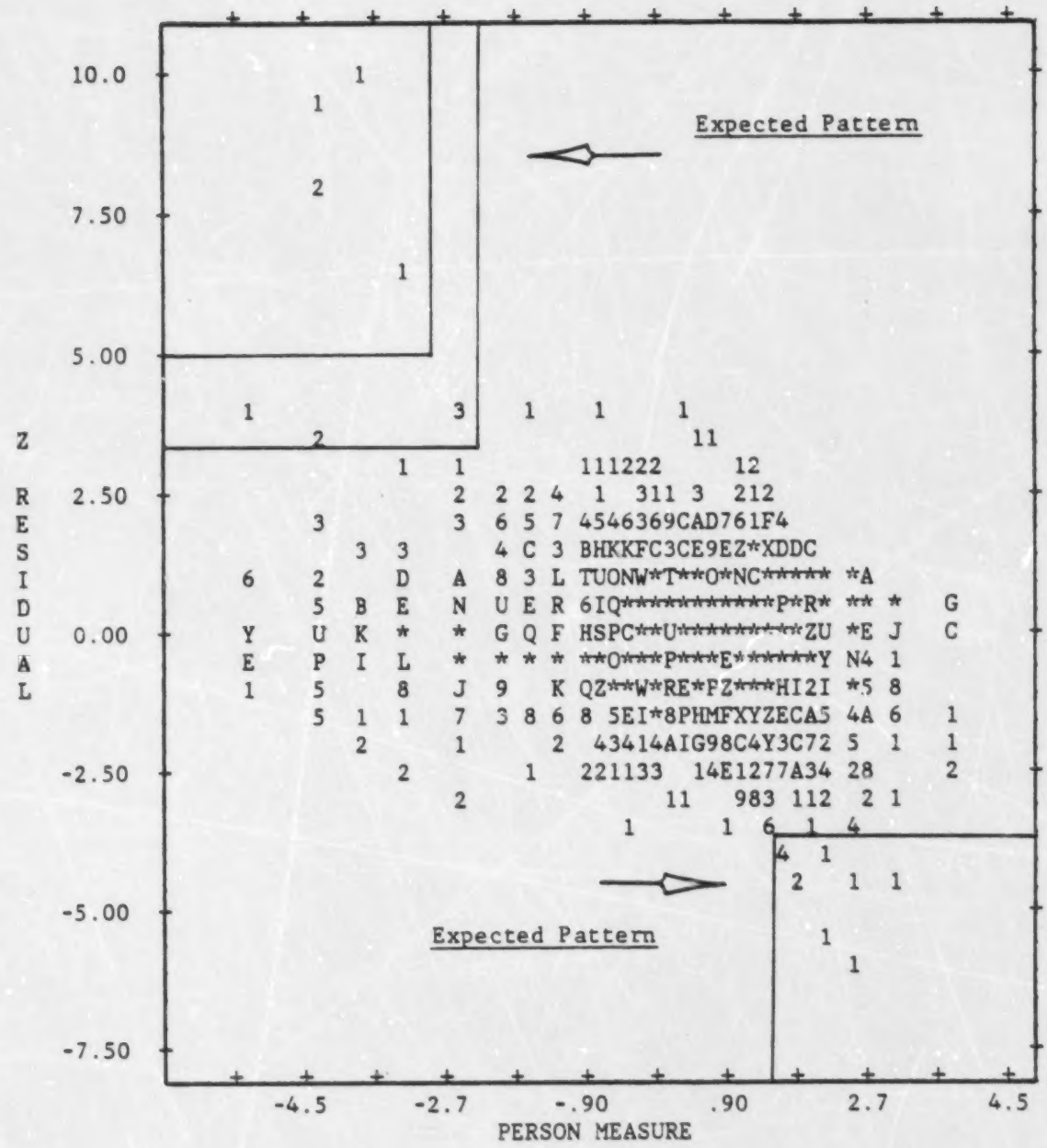


Figure 25--Residuals versus person measures:
Tailored data-first set

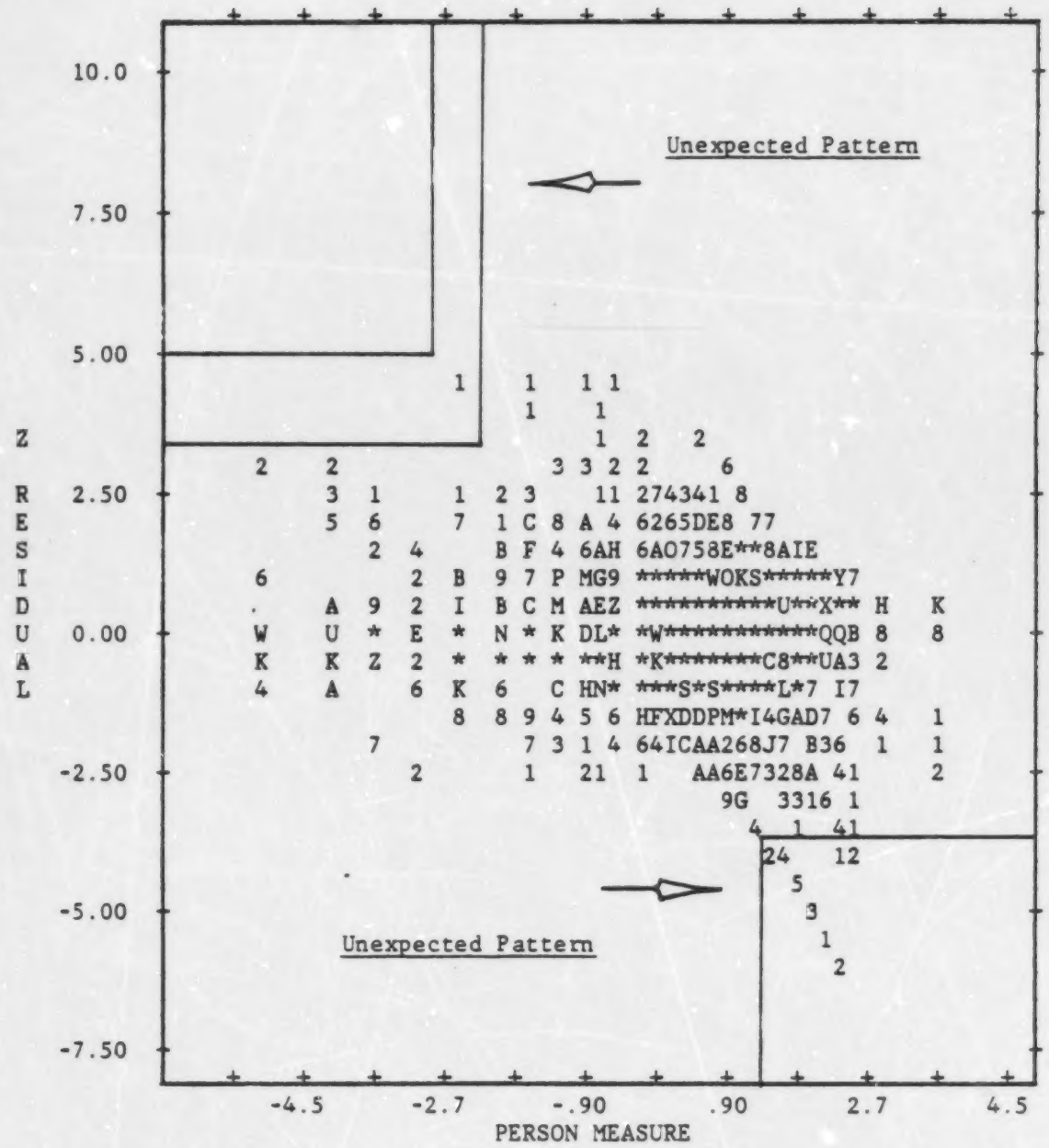


Figure 26--Residuals versus person measures:
Observed data

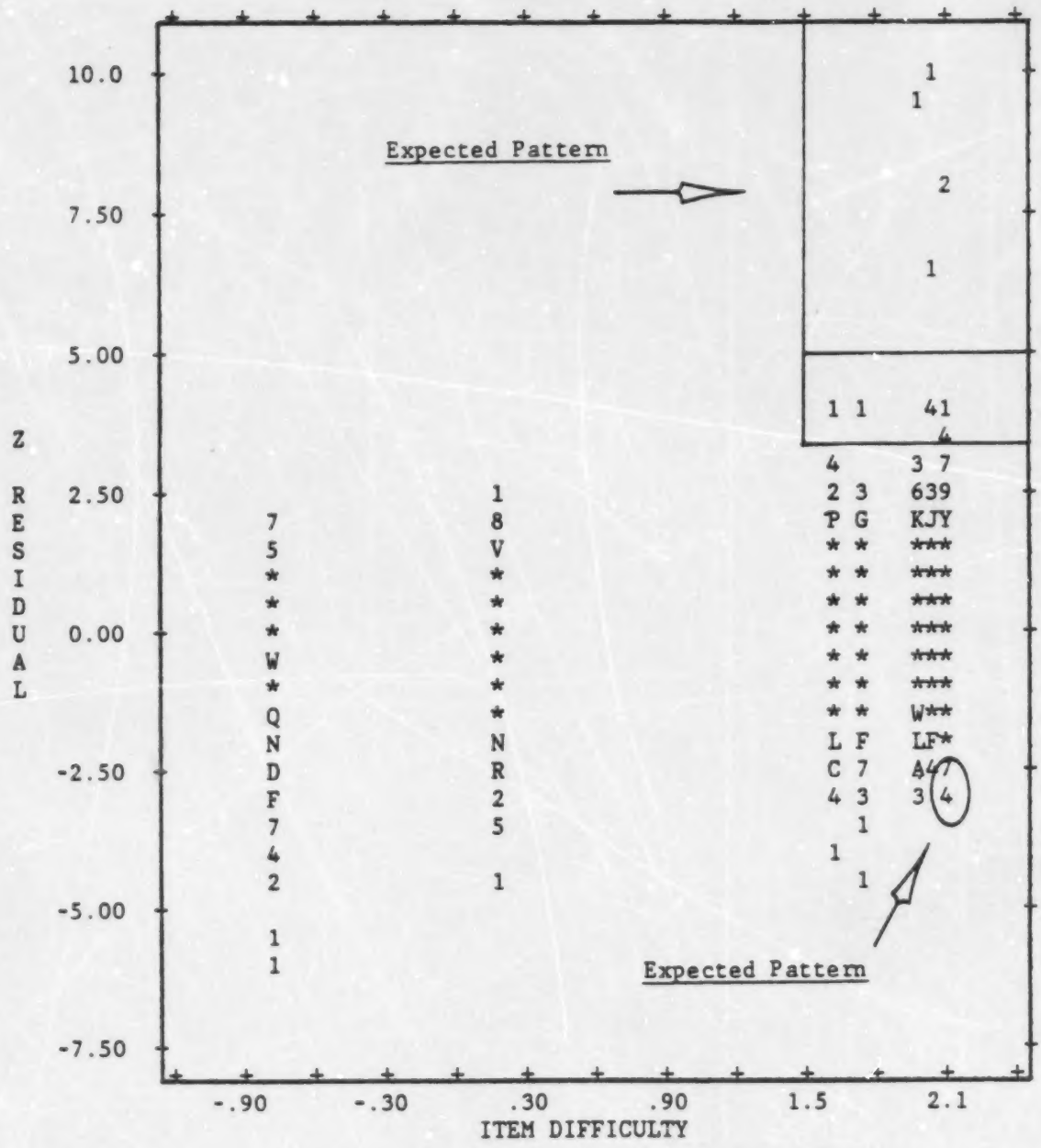


Figure 27--Residuals versus item difficulties:
Tailored data-first set

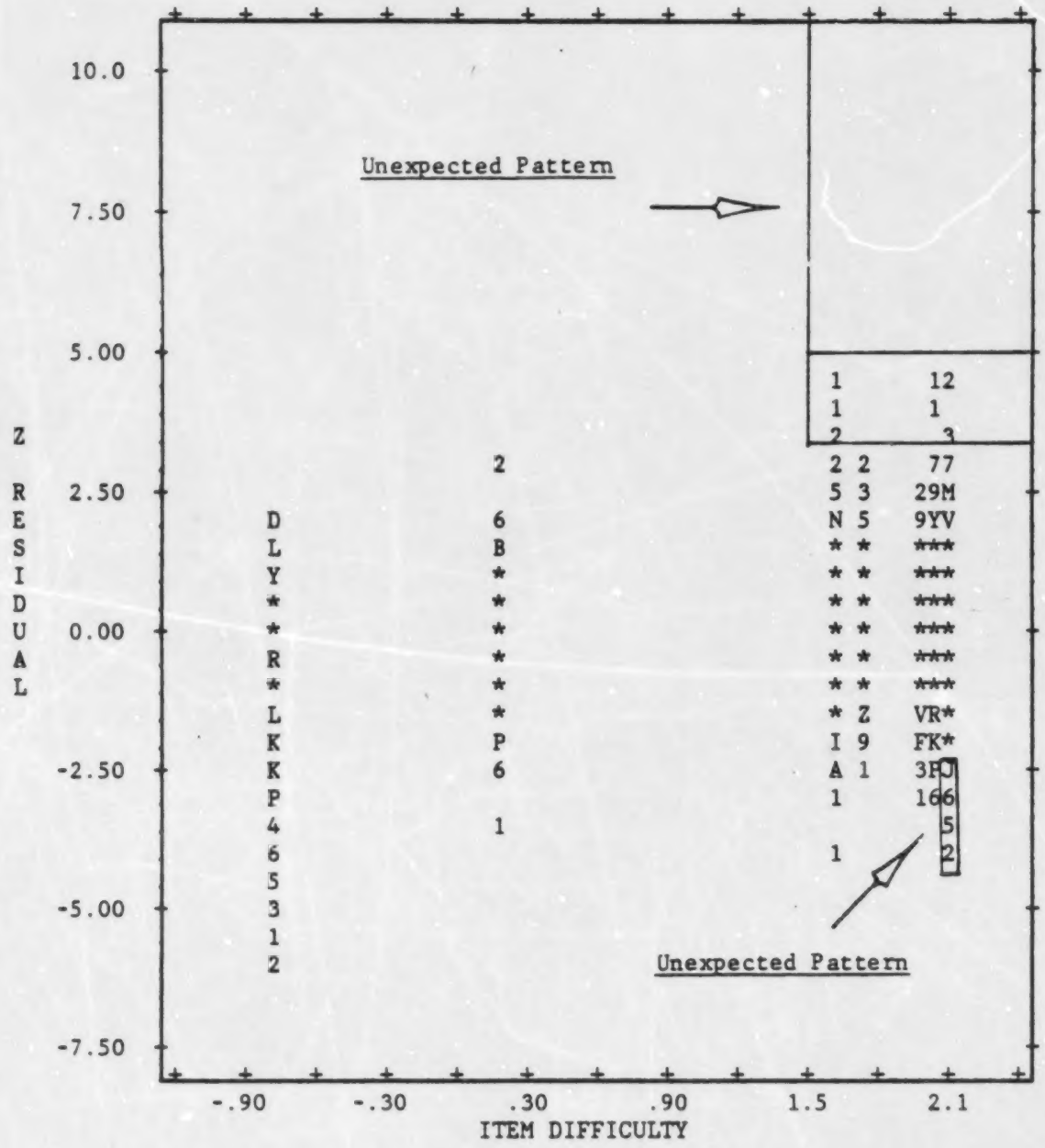


Figure 28 --Residuals versus item difficulties: Observed data

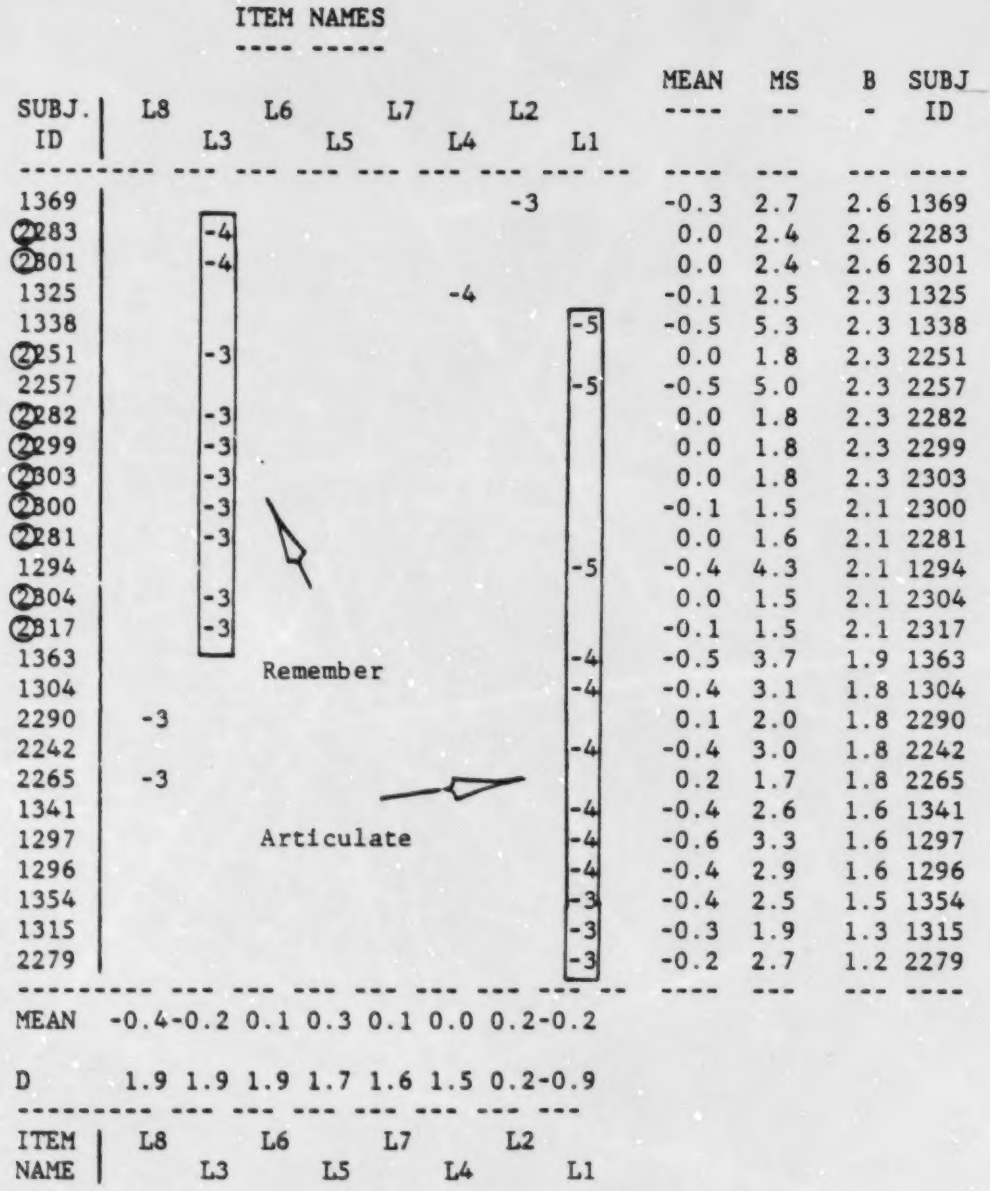


Figure 29--Sorted, truncated, standardized residuals (z<=-3.0), for five year old children: Observed data

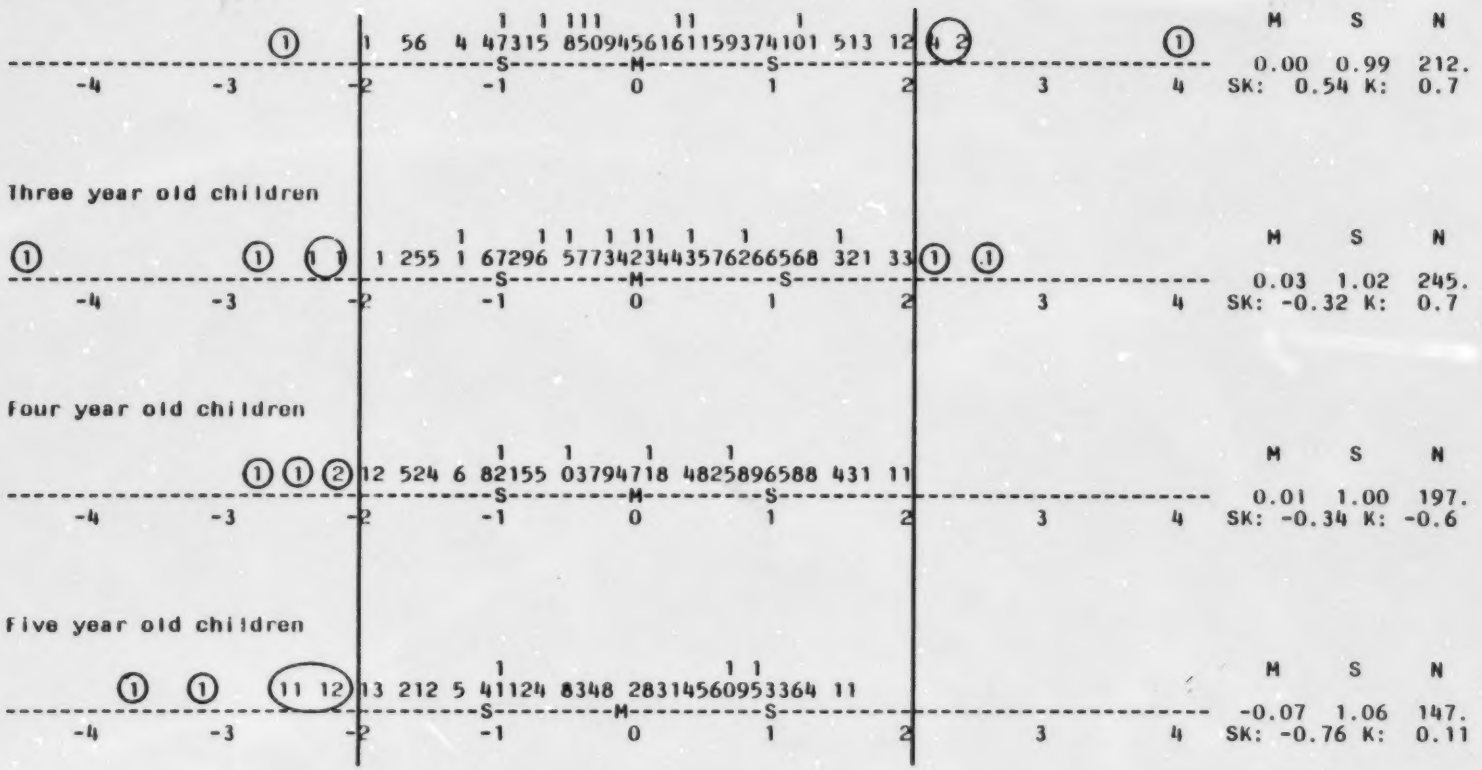


Figure 30 --Residuals on L5, "name action", by age group:
Tailored data-first set (d = 1.71)

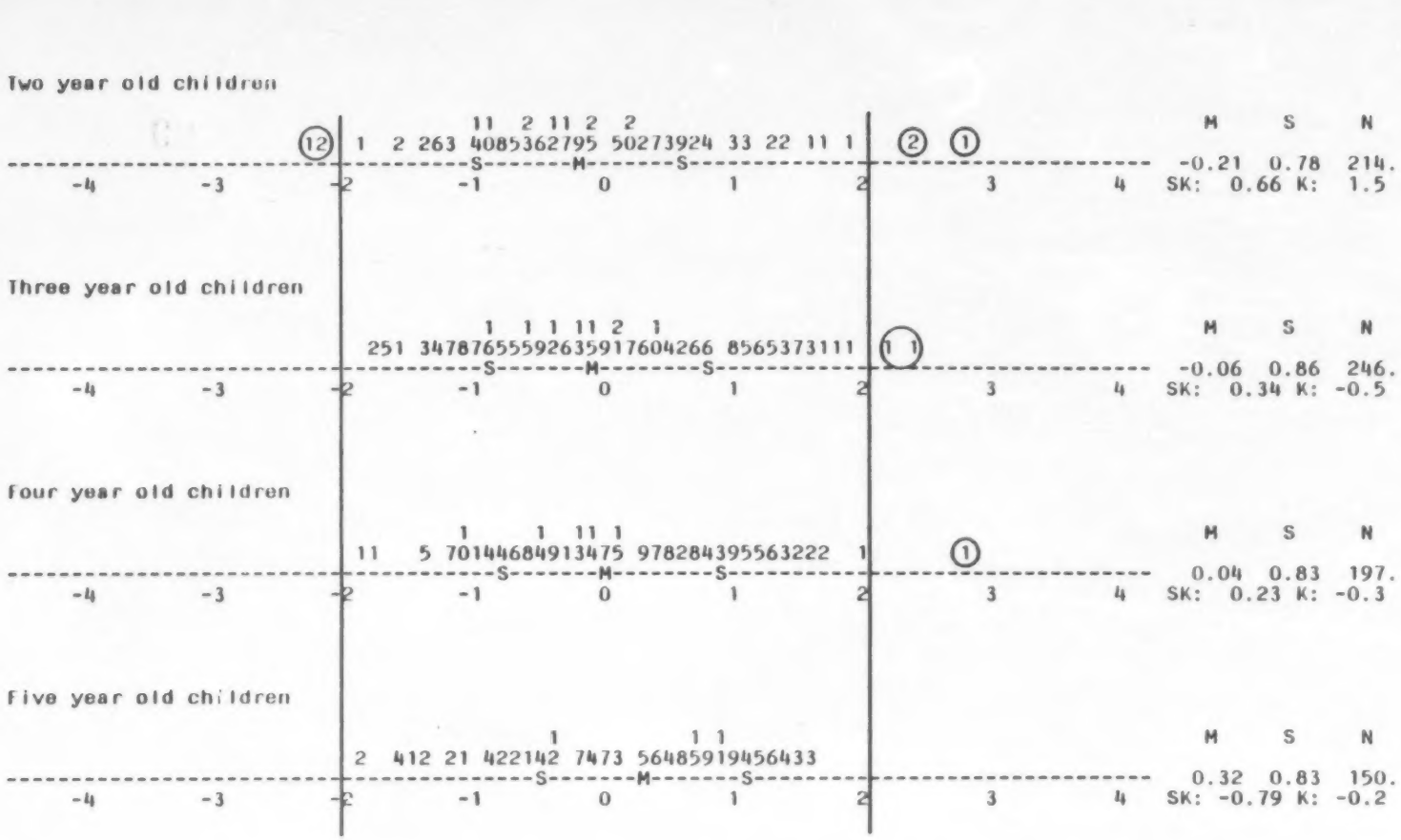


Figure 31--Residuals on L5, "name action", by age group:
Observed data (d = 1.68)